Non-Linear Feature Selection for Biomedical Data Interpretation

POLITECNICO DI TORINO

Ph.D. SCHOOL
Ph.D. Course in Biomedical Engineering – XXVI Cycle

Ph.D. Thesis

Non-Linear Feature Selection for Biomedical Data Interpretation

Samanta Rosati

Supervisor
prof. Gabriella Balestra

Coordinator of the Ph.D. Course
prof. Cristina Bignardi

February 2014

Abstract

The rapid development in computing-related disciplines and technologies provides large amounts of data and variables. *Knowledge Discovery in Databases* (KDD) is a new research field aiming to provide tools and methods for "identifying valid, novel, potential useful, and ultimately understandable patterns in data". KDD is a process that starts from the analysis of the application domain and the identification of the target data. Then, it proceeds with the improvement of the data quality by dimensionality reduction and finally, it applies data mining methods for obtaining new information and knowledge.

The dimensionality reduction is one of the most important step in KDD for the successfully discovery of new knowledge. In fact, in most situations several variables collected in datasets are irrelevant, redundant or could be source of noise for the following analysis phases. Automated methods for *Feature Selection* (FS) allow highlighting only those aspects really relevant in data, deleting all redundant and irrelevant variables. This aspect is important, not only for *Data Mining* (DM) purposes, but also for improving the performances of signal processing algorithms.

In the biomedical field the application of such techniques results very complicated because medical datasets are often characterized by a small number of objects (usually patients involved in the study) associated with a very large number of variables extracted from medical analysis and classes that are usually unbalanced. Moreover, medical data are frequently affected from incompleteness and uncertainty derived from measurement or human errors.

The aim of this thesis is to provide a new methodology for analyzing biomedical data. This approach can be successfully used both for extracting useful

knowledge from real datasets and for improving signal and image processing techniques. In the whole KDD process, we focused on the application of non-linear and automated methods for FS, essentially based on Rough Set Theory (RST). RST is a powerful methodology that does not require any a-priori information or model assumption about data, but it uses only knowledge directly derivable from the given data. Moreover, RST is able to model imperfect and incomplete knowledge, which usually characterizes medical datasets.

The first application is related to the characterization of cerebral hemodynamics of migraine sufferers. A dataset of 26 parameters was built starting from the Near-InfraRed Spectroscopy (NIRS) signals measured on 65 migraineurs and 15 healthy subjects. The QuickReduct Algorithm (QRA), an automated and non-linear FS method based on RST, was applied and compared with the conventional ANOVA analysis. The results show that the variables selected by the QRA, apart from allowing classifying the subjects more accurately than ANOVA ones, were actually related to physiological aspects connected with the pathology.

The second application deals with the characterization of diabetic oxygenation patterns during ankle flexo-extensions. The NIRS signals were acquired in 31 diabetic patients and 16 control subjects, before and after training protocols. Starting from a dataset made of 24 variables, the most discriminative feature subsets were selected by using five automated FS algorithms, all based on RST. A good discriminative power is obtained for all subsets, together with very useful information for the assessment of diabetic peripheral vascular impairment.

Finally, combining these techniques of automatic FS with knowledge-based systems and the traditional image processing methods, a tool for the automatic segmentation of different kinds of ultrasound carotid images is proposed. Several methods have been presented in literature for the automated analysis of US images, whose performances, however, are highly sensitive to the variability introduced by noise, different morphology of the vessel and the presence of disease (plaques). In this study, we started identifying four different classes of pixels, according to their physiological meaning: lumen, intima-media complex, adventitia and noisy lumen. Then, FS by QRA was performed on a dataset made of 600 pixels per class characterized by 211 features. For each single pixel, other than its intensity, we considered as features different parameters essentially based on the intensity of the pixels around it and belonging to two categories: statistical moments estimates and texture features. This process led to the selection of 12 variables that were used for classifying each pixel in one of

the identified classes by means of three Feed-Forward Neural Networks (FFNNs) used in parallel. When all pixels in the region of interest were classified, it was therefore possible to automatically identify the lumen-intima (LI) and media-adventitia (MA) interfaces of the carotid wall. The results are very encouraging: the profiles identified by our tool are comparable with those drawn manually by human operators, also in the presence of noise or plaques in the image. However, we further improved the segmentation performance obtained with this tool, developing a dual-snake system that evolves by means of a Fuzzy Inference System (FIS). This FIS takes as input seven variables related both to internal and external forces connected with the snake definition and, by a set of 26 rules, gives as output the movements of the LI and MA profiles on the image. By this approach, we overcome the main limitations of the classical snake methodology, which needs the user initialization of the snakes and the choice of the correct parameters for the snake evolution. On the contrary, the proposed fuzzy-snake system is completely automatic and independent from the user because the two snakes start their evolution from the LI and MA interfaced identified with the pixel classification tool. Moreover, no parameters are required for the force equilibrium, thanks to the FIS capability to balance the different contributions. We validated our system on 180 pathological and non-pathological images and the results showed very good performances compared with the manual segmentation and the classical semi-automatic techniques, improving the results obtained with the pixel classification tool.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

1CM	1st classification method
2CM	2nd classification method
APA	Adapted Physical Activity
ACO	Ant Colony Optimization
AntRSAR	Ant Rough Set Attribute Reduction
AIS	Artificial Immune System
AI	Artificial Intelligence
ANN	Artificial Neural Network
ANN	Artificial Neuron
BH	Breath-Holding
BHI	Breath-Holding Index
CVD	CardioVascular Disease
CA	Carotid Artery
CS	Classical Snake
CI	Computational Intelligence
Contr	Controls
DM	Data Mining
DR	Dimensionality Reduction
EBR	Entropy-Based Reduction
EC	Evolutionary Computation
ADF	Far Adventitia
FS	Feature Selection
FFNN	Feed-Forward Neural Network
FW	Fit Walking
FIS	Fuzzy Inference System
FSys	Fuzzy System
GA	Genetic Algorithm
GenRSAR	Genetic Rough Set Attribute Reduction
GLDM	Gray Level Difference Method
GLRLM	Gray Level Run Length Matrix
GLCM	Grey Level Co-occurrence Matrix
GT	Ground Truth
HYP	HYPerventilation
IQRA	Improved QuickReduct Algorithm
IMT	Intima-Media Thickness
KDD	Knowledge Discovery in Databases
LF	Low Frequency

LI	Lumen-Intima
MA	Media-Adventitia
MF	Membership Function
MwA	Migraine with Aura
MwoA	Migraine without Aura
NIRS	Near-InfraRed Spectroscopy
O_2Hb	Oxygenated Hemoglobin
PSO	Particle Swarm Optimization
PxCl	Pixel Classification tool
PDM	Polyline Distance Metric
QRA	QuickReduct Algorithm
RBFNN	Radial Basis Function Neural Network
HHb	Reduced Hemoglobin
RST	Rough Set Theory
SOM	Self-Organized Map
SNR	Signal-to-Noise Ratio
SGLDM	Spatial Gray Level Dependence Matrix
SCF	Squared Coherence Function
SVM	Support Vector Machine
SI	Swarm Intelligence
TOI	Tissue Oxygenation Index
VPRS	Variable Precision Rough Set
VLF	Very Low Frequency

Chapter 1: Introduction

The rapid development, during the last decades, in computing-related disciplines and technologies produced a large increase in the amount of data that can be generated, collected, processed, analyzed and stored. As consequence of this situation, the common techniques for examining datasets, mainly based on the manual analysis and interpretation of data, are not applicable, as well as very expensive from time and cost point of view.

Therefore, simultaneously, a new research field emerged, related to the development of new techniques for automated data analysis and extraction of knowledge from large datasets. The term *Knowledge Discovery in Databases* (KDD) identifies the process for "extracting high-level knowledge from low-level data in the context of large datasets" [1]. It is based on statistics and intelligent and automatic methods helping researcher and practitioners to find *patterns* in data, that could be models, structures or, in general, any high-level description of data [1]. KDD is a non-trivial process that starts from an exhaustive analysis of the application domain and the identification of the target data in terms of objects and variables. Then, it proceeds with the improvement of the data quality in order to remove noise and irrelevant or redundant information. Finally, *Data Mining* (DM) methods are properly chosen and applied for obtaining information and, if correctly interpreted in the context of analysis, new knowledge.

Unfortunately, most of the methodologies and methods actually proposed in KDD field focus mainly on the DM step, and two definitions are often confused and used as synonyms. DM is only one step of the KDD process, that concerns the application of specific algorithms and methods for the extraction

of patterns or models from data, without any further pre or post-elaboration [2]. It is clear that, to successfully discover knowledge from large datasets, all steps of the KDD process have to be analyzed and correctly performed. In particular, data reduction is fundamental for the following analyses but, at the same time, it often represents a source of significant data loss [3].

In this context, the dimensionality reduction of multivariate data is a powerful method for highlighting only those aspects really relevant in data, reducing the measurement and storage requirements. Moreover, a too large number of features does not necessarily allow increasing the classification accuracy: several attributes may introduce some kind of noise which decreases the classification performances [3]. In particular, *Feature Selection* (FS) is a procedure allowing the dimensionality reduction of large datasets by choosing the minimal subset of features from the initial set of variables containing the same amount of information than the original set. Those attributes that do not affect the classification performance (*irrelevant*) or highly correlated with other variables and interchangeable with them (*redundant*) are automatically deleted from the dataset during the FS, so that the remaining features result the most significant for the system description. Moreover, FS preserves the meaning of the features, facilitating the phase of knowledge interpretation, unlike *feature construction* methods (such as the Principal Component Analysis or the Linear Discriminant Analysis) which perform a transformation of the original space, making the final interpretation more difficult.

As an exhaustive search of the best feature subset results inapplicable in most of real situations, the research in the FS field proposed, in the last years, several methods for finding the best variable subset based on a heuristic search [4, 5]. Heuristics identifies a wide class of algorithms used in order to solve optimization problems, reducing time and computational costs by addressing the solution search toward a high-quality space of admissible solutions.

The application of such techniques for biomedical data analysis is poorly dealt in literature, because of many problems related to their management. Firstly, medical datasets are very often made up of few elements, normally representing patients involved in a study, characterized by a very high number of variables extrapolated from the huge set of medical analysis and examinations today available. Moreover, medical data are usually affected by high imbalance in the number of elements for each considered class. For example, in studies dealing with specific pathologies, usually a higher number of patients are available while healthy subjects, needed as controls, are poorly represented. Finally, this kind of datasets frequently suffers from incompleteness and uncertainty

derived from measurement or human errors. Due to these characteristics, analyzing biomedical data is not a trivial task. Building models and classifiers able to manage the incomplete information contained in data, often hidden by the very large number of features, requires expertise and it is accomplished by means of sometimes long processes.

The aim of this study is to provide a new effective and efficient methodology for successfully managing biomedical data, in order to facilitate the extraction of new useful knowledge from real datasets and improve signal and image processing techniques. In this process, special attention is given to the phase of FS, conducted by using new advanced tools based on the Rough Set Theory (RST). RST is a new methodology born as extension of the Fuzzy Logic that has been applied in the FS field with very satisfactory results [6, 7]. It enables to manage incomplete and imperfect knowledge encapsulated in the real data without any assumption about data model.

Chapter 2 contains a brief introduction to the KDD process and an overview to the FS problem and the techniques currently available in this field.

Chapter 3 is mainly focused on the description of the materials and methods used in this thesis, with an overall description of the computational intelligence context and the RST foundations. Particular attention is given to the supervised FS algorithms employed for dimensionality reduction and to the intelligent systems used for DM.

In Chapter 4, the first application of these techniques to a small and unbalanced medical dataset is presented. Specifically, the KDD process is used to characterize the cerebral hemodynamics of migraine sufferers, basing on Near-InfraRed Spectroscopy (NIRS) signals. A dataset made of 26 parameters was built involving 65 pathological and 15 healthy subjects. Two FS methods were applied: the QuickReduct Algorithm (QRA), an automated and non-linear FS method based on RST, and the conventional ANOVA analysis. The performances of the two selected subsets were compared by means of Feed-Forward Neural Networks (FFNNs), in order to assess their capability in the correct identification of the subject's class.

The second application of these techniques is presented in Chapter 5, dealing with the characterization of diabetic oxygenation patterns during ankle flexo-extensions. The NIRS signals were acquired in 31 diabetic patients and 16 control subjects, before and after training protocols and 24 variables were extracted for each of them. The most discriminative variables were selected by using five automated FS algorithms, all based on RST and the results are again

compared in terms of correct classification of the subjects, by using FFNNs and Support Vector Machines (SVMs).

Finally, combining these techniques for automatic FS with knowledge-based systems and traditional image processing methods, an innovative tool for the completely automatic segmentation of ultrasound carotid images is described in Chapter 6. This complex application was realized by classifying all image pixels in the region of interest using a set of 12 features related to the pixel intensity and selected by QRA. The profiles of the carotid layers obtained with this intelligent tool, identifying the boundary between lumen and intima layers and between media and adventitia layers, were further elaborated by Fuzzy Logic, as reported in Chapter 7. A Fuzzy Inference System (FIS) was constructed that drives the evolution of two active parametric contours (or snakes) for recognizing the interfaces of interest. In this case, the developed system is completely automated, as it does not require any profile initialization and parameter tuning.

Chapter 8 draws the conclusions, highlighting the innovative aspects of this work.

References

1. Fayyad, U., G. Piatetsky-shapiro, and P. Smyth, *From Data Mining to Knowledge Discovery in Databases.* AI Magazine, 1996. **17**: p. 37-54.
2. Fayyad, U.M., G. Piatetsky-Shapiro, and P. Smyth, *From data mining to knowledge discovery: an overview*, in *Advances in knowledge discovery and data mining*, M.F. Usama, et al., Editors. 1996, American Association for Artificial Intelligence. p. 1-34.
3. Jensen, R. and Q. Shen, *Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches*. 2008, Hoboken, NJ: Wiley-IEEE Press. 340.
4. Saeys, Y., I. Inza, and P. Larranaga, *A review of feature selection techniques in bioinformatics.* Bioinformatics, 2007. **23**(19): p. 2507-17.
5. Dash, M. and H. Liu, *Feature selection for classification.* Intelligent Data Analysis, 1997. **1**(1-4): p. 131-156.
6. Wang, J., et al., *Rough set and scatter search metaheuristic based feature selection for credit scoring.* Expert Systems with Applications, 2012. **39**(6): p. 6123-6128.
7. Geng, Z. and Q. Zhu, *Rough set-based heuristic hybrid recognizer and its application in fault diagnosis.* Expert Systems with Applications, 2009. **36**(2, Part 2): p. 2711-2718.

Chapter 2: Knowledge Discovery in Databases and Feature Selection

2.1 Knowledge Discovery in Databases

Since 1990s, the rapid development in computing-related disciplines produced a rapid increase of the amount of data that can be generated, measured and acquired in the field of science, business and government. Moreover, the progresses in data storage and database technologies allowed collecting and storing huge quantities of data rapidly, efficiently and inexpensively. The growth of databases was both in terms of number of objects (or records) included, and in terms of number of variables (or attributes) characterizing each element [1].

In this context, the traditional approaches of data analysis, based essentially on the manual examination of datasets, became impracticable, as it is unavoidably time and cost expensive and highly subjective. These considerations led researchers and practitioners to create, during the last two decades, new techniques for automated data analysis and consequently for extracting useful information from large volume of data. These new tools are included in the emerging field of KDD, born with the aim to provide intelligent and automatic methods for "extracting high-level knowledge from low-level data in the context of large datasets" [1].

From its first introduction during the first KDD workshop in 1989 [2], several definitions were proposed for KDD, but the most widely used was given by Frawley *et al.* [3]:

> "*knowledge discovery in databases is a non-trivial process of identifying valid, novel, potential useful, and ultimately understandable patterns in data*".

This sentence summarizes the most important aspects in KDD. Firstly, KDD research focuses on *data* (or observations) describing different aspects of the system of interest [4]. The goal of KDD is to find *patterns* that could be models, structures or, in general, any high-level description of data [1]. Then, KDD is a *process* involving many steps that have to be performed iteratively and with the user interaction. Finally, the term *non-trivial* highlights that obtaining high-quality information from data is not an easy task. Obstacles come mainly from the fact that data are rarely collected for the mining of knowledge but rather they are a byproduct of other processes. As consequence, essential variables are often missing during the data collection. Moreover, very often the entire population is not involved in the study, but only limited samples with specific characteristics are available [5].

Nowadays, the term KDD is often confused with DM and, in most scientific reports, the two definitions are used as synonyms. Although from a systematic literature review it emerges that systems labeled as KDD systems often propose no more than a DM technique [6], there is a clear difference between the two definitions. DM represents only a single step of the whole KDD process, dealing with the application of specific algorithms and methods for the extraction of patterns or models from data, without any further elaboration [5]. On the opposite, KDD identify the complete procedure of discovering information, requiring also a-priori knowledge about the domain of interest and final interpretation of the DM results in the context of application.

2.1.1 KDD Process

As stated in the definition given by Frawley *et al.* [3], KDD is a process and, as such, it is composed by several steps illustrated in Figure 2.1 and described below [1].

- ✓ *Understanding of the Application Domain:* this first step, concerning a deep analysis of the context of application and of the relevant a-priori knowledge, is required in order to identify the goal of the KDD process from the user point of view.

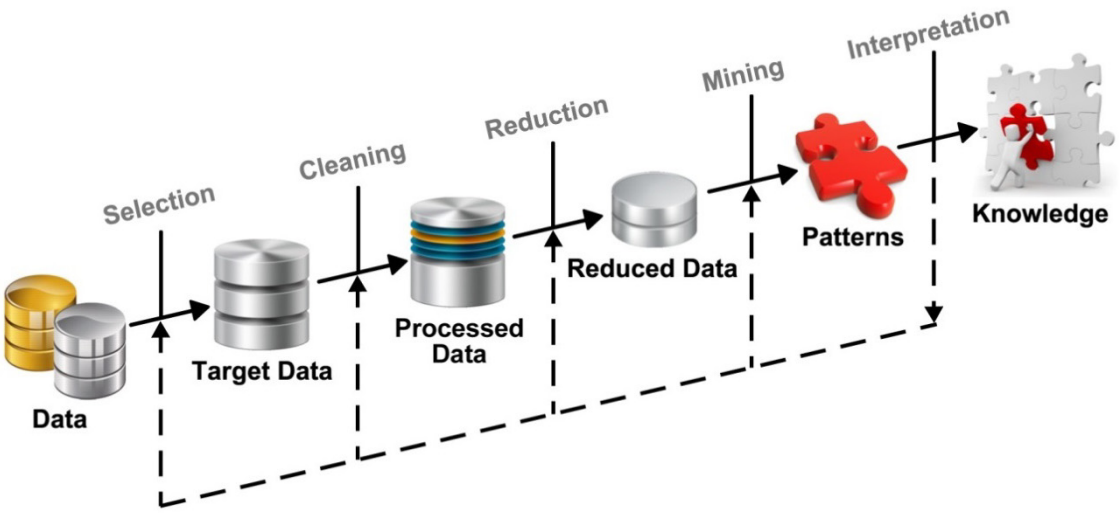

Figure 2.1 An overview of the main steps of the KDD process [7].

- ✓ *Data Selection:* in this step a proper target dataset for KDD is identified. It could be created through the selection of a subset of variables or samples from a larger dataset or obtained by means of the union of several existing datasets.
- ✓ *Data Cleaning:* this preprocessing operation includes all procedures aiming to improve the quality of data and consequently to facilitate the discovery of knowledge, such as noise removal, managing of missing data, attribute discretization, …
- ✓ *Data Reduction:* the goal of this phase is to remove irrelevant or redundant variables from the dataset in order to preserve only those features useful for representing the data. Moreover, in case of large datasets, this procedure allows the reduction of measurement and storage requirements, and elaboration time.
- ✓ *Data Mining:* the choice of the appropriate DM method, in terms of model and parameters to be used, have to consider the goal of the KDD process (classification, regression, clustering, …) and the values contained in the dataset (categorical, numerical, continuous, discrete,…).
- ✓ *Interpretation:* once the knowledge is extracted from data, it requires to be interpreted and evaluated regarding its validity and usefulness in resolving the starting problem. If the obtained results are not interesting according to the user opinion, one of the previous phases could be repeated.

- ✓ *Use of the discovered knowledge:* the discovered knowledge can be used directly, incorporate into other systems or simply documented and reported to the interested parties.

Most of the research actually proposed in KDD field focuses mainly on the DM step, but the other phases are not less important for the successful application of KDD. In particular, data reduction is a fundamental step for the following analyses but, at the same time, it often represents a source of significant data loss [7]. *Dimensionality Reduction* (DR) of multivariate data represents a set of powerful methods for automatically deleting those attributes that are not predictive of the final state of the system (*irrelevant*) or highly correlated with other variables and interchangeable with them (*redundant*), and for highlighting the features relevant for the system description. Two different techniques are available for performing DR on datasets: *feature construction* and *feature selection* (FS) [8].

All techniques generating a completely new set of features from the original one belong to the first category, in which a transformation of the original space is performed in order to represent data in a more efficient way. This type of procedures, such as Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA), results in a more difficult interpretation because the new features do not correspond to the original ones. Moreover, since the new attributes are obtained as a linear or non-linear combination of all initial ones, all features needed to be always collected in order to generate the new set of variables.

FS algorithms simply select a minimal number of features from the initial set of variables, preserving the meaning of the features. This characteristic, that facilitates the phase of knowledge interpretation, makes FS algorithms more suitable for KDD.

2.2 Feature Selection

The main idea behind a FS method is to reduce the dimensionality of data selecting a minimal number of features from the initial set of variables. The extracted features result the most significant for the system description, keeping intact the amount of information with respect to the original variables. Although it could be expected that the increase of the number of variables leads to a resulting increase of information about the problem, this is not true due to the *course of dimensionality* [9]: an increment in the dimensionality produces an exponential dispersion of the data samples in the high-dimensional space. An ex-

ample of the course of dimensionality is shown in Figure 2.2 for 25 samples generated from a uniform distribution and represented in (a) one dimension, (b) two dimensions and (c) three dimensions. It can be observed that, increasing the dimensionality, samples results more spare. In particular, considering a unit-size box (in red in Figure 2.2), there are 15 samples within the box in one-dimension, 7 samples in two dimensions and only 3 samples in three dimensions. Specifically, the sample density in a M-dimensional space is proportional to $N^{1/M}$, where N is the total number of considered points. This means that, a sample density of 25 in one dimension will require 25^3=15625 points in three dimensions to reach the same density [10].

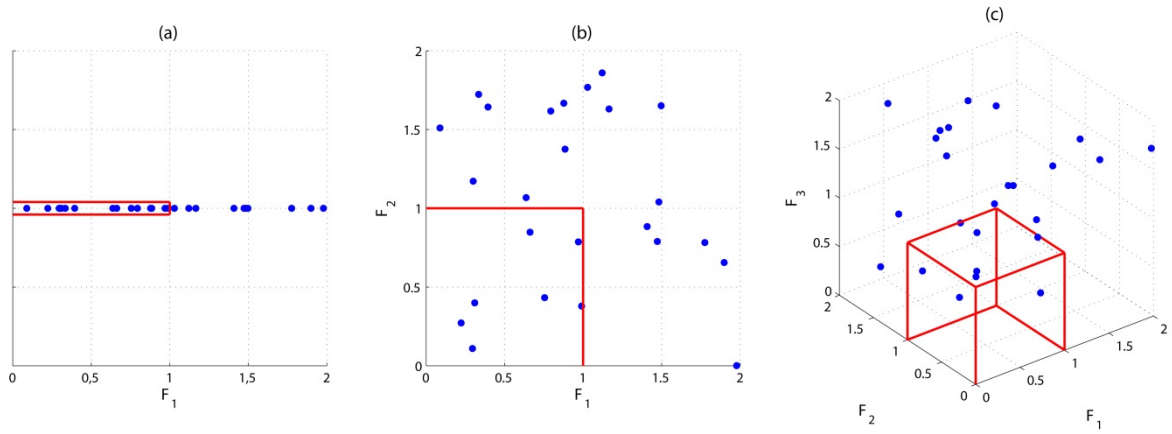

Figure 2.2 Illustration of the course of dimensionality. In this example, 25 samples are generated from a uniform distribution between 0 and 2. (a) Plot in one dimension, (b) plot in two dimensions, (c) plot in three dimensions. In red unit-size boxes are identified: there are 15 samples within the box in (a), 7 samples in (b) and only 3 samples in (c) [10].

Further advantages can be achieved by reducing the dimensionality of problems thanks to the improvement of the inductive learner in terms of learning speed and generalization capacity, the reduction of measurement cost and the better understanding of the problem domain [11]. For these reasons FS algorithms were widely applied in a lot of fields such as image recognition [12, 13], text categorization [14, 15], clustering [16, 17], bioinformatics [18, 19], rule induction [20, 21] and system monitoring [22, 23] (Figure 2.3).

Different approaches can be followed to select the best feature subset. An exhaustive search could be performed exploring the whole space of possible subsets (*brute-force approach*) [24]. However, such a procedure results inapplicable when the number of initial variables is relatively medium-high, which is for most of real applications. In fact, considering N objects characterized by M fea-

tures, the number of subsets that can be obtained as a combination of the original attributes is given by $N!/[M!\,(N-M)!]$.

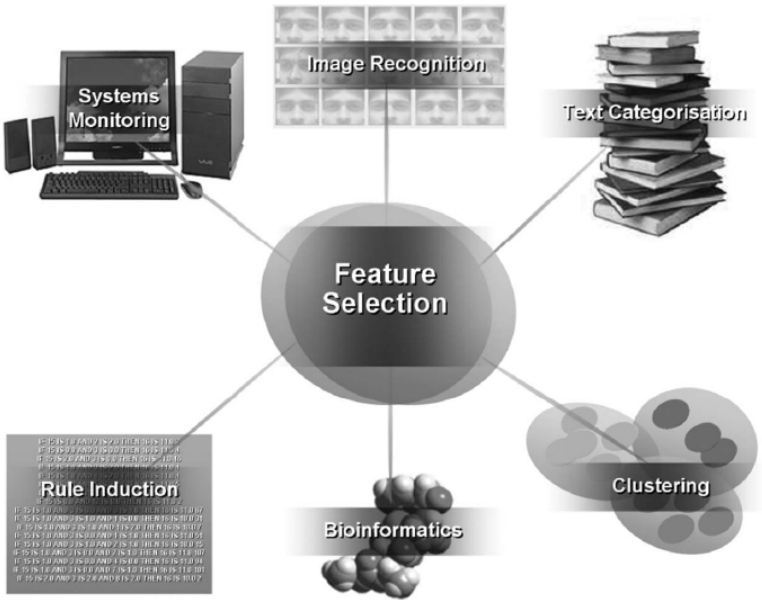

Figure 2.3 Typical FS application areas [7].

These considerations led to the development, during the past years, of several methods for FS based on *heuristic search* [24, 25]. Heuristics identify a wide class of algorithms used to solve optimization problems bypassing the complexity induced by real world applications. Time and computational costs reduction is achieved by addressing the solution search toward a high-quality space of admissible solutions. They are based on the evolution of a heuristic function that provides an estimate of the current solution goodness. Their adaptability to different fields with relatively few modifications makes heuristic algorithms a suitable tool also for FS. A typical FS method based on heuristic search is built as an iterative procedure following these steps [25] (Figure 2.4):

1. *Subset generation*: the feature subset is updated gradually, so that there is an improvement in its accuracy. This phase can start with an empty subset, with a subset containing all features or with a randomly selected subset. In the first two cases, features are iteratively added (*forward selection*) or removed (*backward selection*), whereas in the last case, features are iteratively either added or removed or produced randomly.
2. *Subset evaluation*: an evaluation function is needed to measure the goodness of the current subset. If this value is better than the one of the previous best subset, the current subset becomes the best one.

3. *Stopping criterion:* steps 1 and 2 are iteratively performed until a stopping criterion is reached. Different stopping criteria can be used: a predefined number of features are selected, a predefined number of iterations is reached, the addition (or deletion) of any feature does not produce a better subset, and an optimal subset according to some evaluation function is obtained.
4. *Validation:* different tests can be performed on the selected subset in order to assess its validity in resolving the specific task for which it was constructed.

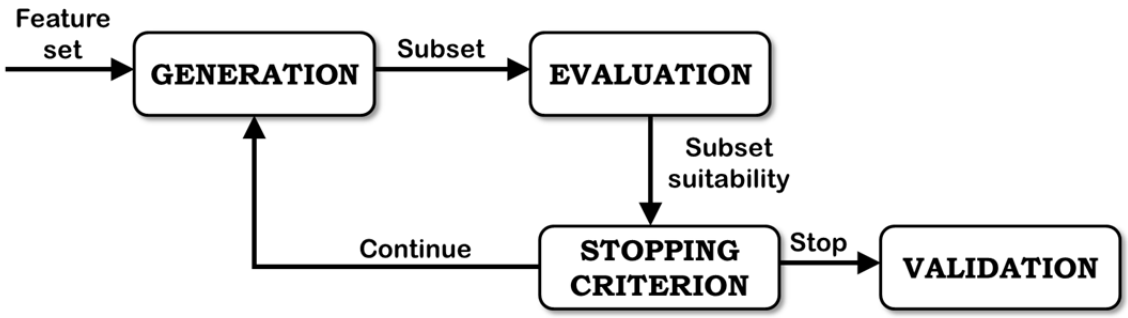

Figure 2.4 Typical FS process based on heuristic search [7].

2.2.1 Feature Search

FS algorithms are commonly organized into three categories according to the relationship between the selection method and the learning algorithm, namely filter, wrapper and embedded methods [8, 11, 26] (Figure 2.5).

Filter methods perform FS independently of the learning algorithm. This means that data are analyzed in order to identify those dimensions that are more relevant for describing their inner structure. Hence, the filter procedure results are not correlated with the learning task and computationally easy but, as each features is considered independently, groups of features having strong discriminatory power are ignored.

In *wrapper methods,* the feature subset is assessed by estimating the goodness in the learning task, allowing to reach better performances than with filter methods and to capture also feature dependencies. Obviously, this kind of algorithms is computationally very expensive and strongly influenced by the chosen learning procedure (the feature subset is optimized for that specific algorithm).

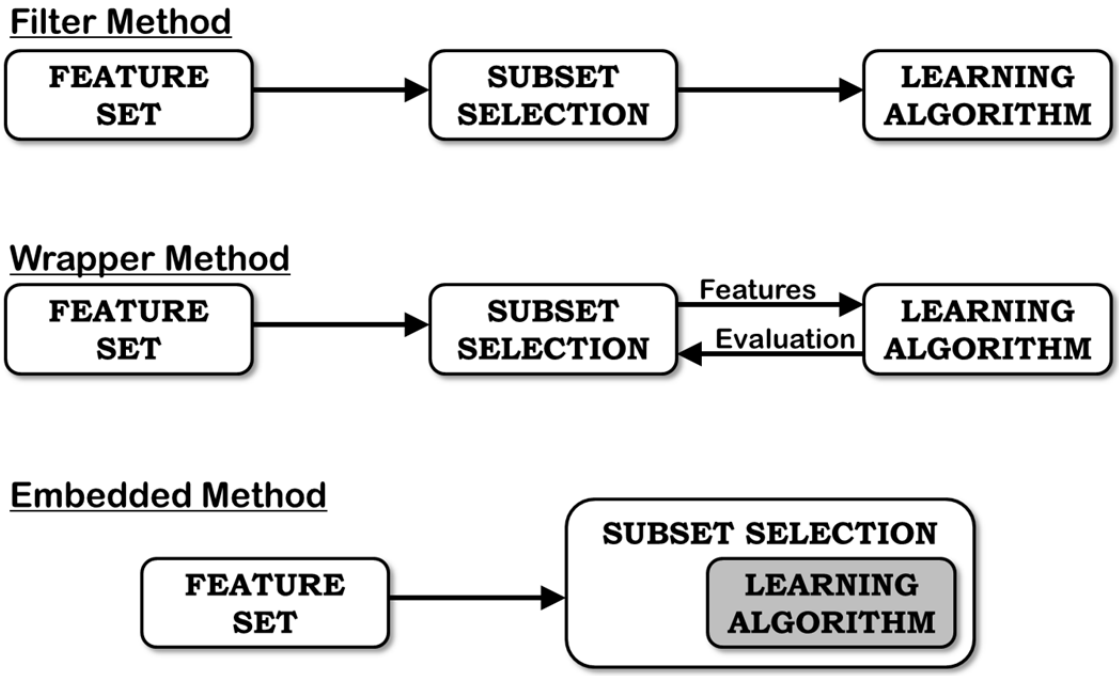

Figure 2.5 Schematic representation of the different approaches used for FS [7].

In *embedded methods,* the FS is performed during the training phase of the learning algorithm, allowing to reach better performances than with filter methods with computational cost lower than wrapper methods. Also in this case, there is a high dependency of the selected subset from the learning procedure but the interaction among features can be captured.

Errore. L'autoriferimento non è valido per un segnalibro. reports an overview of the main advantages and disadvantages for the different typologies of FS algorithms.

2.2.2 Subset Evaluation

In order to evaluate the goodness of a selected subset of features a proper function is required, that is able to measure the discriminant ability of the variable subset in grouping the objects. Several metrics were proposed to test the relevance and non-redundancy of a subset. These can be divided into five categories [25]:

- ✓ *Distance:* a feature X is preferable with respect to a feature Y if it induces a greater separability (in terms of distance) in the object classes while the two variables are indistinguishable if the class distance does not change. An example of this evaluation function is the Euclidean distance.

Table 2.I Overview of the main advantages and disadvantages for the different typologies of FS algorithms.

Method	Advantages	Disadvantages
Filter	Independence from the learning algorithm Low computational cost Fast Good generalization ability	No interaction with the learning algorithm
Wrapper	Interaction with the learning algorithm Captures feature dependencies	Computationally expensive Risk of over fitting Selection dependent from the learning algorithm
Embedded	Interaction with the learning algorithm Low computational cost Captures feature dependencies	Selection dependent from the learning algorithm

- ✓ *Information:* it measures the information gain due to a feature. This means that a variable X is better than a variable Y if it produces a higher information gain. Otherwise, if the information gain is zero, the feature is irrelevant. The entropy measure belongs to this class of functions.
- ✓ *Dependence or Correlation:* this is a measure of the ability of a variable to predict the value of another. For example in the supervised FS procedures (§2.2.3), a feature X is preferable to a feature Y if X has a higher correlation with the classification than Y.
- ✓ *Consistency:* according to this measure, the feature subset is chosen as the one with minimum number of variables and a degree of inconsistency lower than a threshold sets by user. In supervised context (§2.2.3), inconsistency is defined as two instances having the same feature values but different class labels [27].
- ✓ *Classification Error Rate:* with this approach a feature subset is selected basing on the accuracy level it produces in the classification task. This means that this kind of metric can be applied only for supervised learning (§2.2.3) and, consequently, a high correlation with the chosen classifier is produced.

The comparison of the five categories of evaluation functions in terms of *generality (*the subset performances in the learning task using different learning algorithms), *time complexity (*the time required for the subset selection), and the *accuracy* in the class prediction is reported in Table 2.II.

Table 2.II Comparison of different categories of evaluation function in terms of generality, time complexity and accuracy [25].

Evaluation Function	Generality	Time Complexity	Accuracy
Distance Measure	Yes	Low	-
Information Measure	Yes	Low	-
Dependence Measure	Yes	Low	-
Consistency Measure	Yes	Moderate	-
Classification Error Rate	No	High	Very High

2.2.3 Supervised FS

Supervised FS aims at selecting the minimum set of features necessary to discriminate which class the objects belong to. This means that the dataset used for FS must contains, not only the features characterizing each element, but also the information about their class.

Because an exhaustive analysis of all FS algorithms proposed in literature is beyond the aim of this work, in the next paragraphs there are only an overview of the most important methods developed in this field, divided according to the employed search strategy.

2.2.3.1 Filter Methods

Several algorithms were proposed during the last years in this category. Almost all of them are based on heuristic search and employ different criteria for the subset evaluation.

Heuristic algorithms based on distance measure are very common in literature [28-31]. One of the most popular, *Relief* [28], associates a weight to each feature based on the distance between an object randomly chosen from the dataset and the nearest neighbors in the same class and in the opposite class. At

the end of a fixed number of iterations, only the features with weight higher than a threshold set by user are selected.

Information gain is used in *minimum Redundancy Maximum Relevancy* (mRMR) [32],employing the mutual information between an attribute and the classification feature in order to select those attributes with maximum relevancy with the class and minimum redundancy.

In *Correlation-based Feature Selection* (CFS) [33]the correlation measure is used to find the feature subset highly correlated with the object classes (relevant features) and without any redundant variable. Seth and Principe [34] proposed an extension of mRMR in which the mutual information is substituted by the dependence measure to assess relevance and redundancy.

An example of FS algorithm with evaluation functions based on consistency is proposed by Almuallim and Dietterich [35]. In this method, called *Focus,* all subsets with a certain number of features are generated, starting from one variable, until a consistent subset is found. This kind of approach results very sensitive to noise and inapplicable to datasets with medium-high dimensions for the complete generation procedure it performs.

Only few approaches proposed in literature are based on the random generation of the feature subsets. The most famous method belonging to this group is the *Las Vegas Filter* (LVF) [36] in which, starting from the entire set of features considered as the best subset, a random subset is extracted for a certain number of iterations. At each iteration, the current subset is compared with the best one: if it has a number of features lower than the current best and an inconsistency rate minor or equal to a threshold set by user, it is considered the new best subset. Also in this case, the calculation of the consistency can take a long time for medium-high datasets.

All algorithms previously cited generate feature subsets by incrementally adding or removing individual features. Recently some approaches based on feature grouping have been proposed [37-39] which evaluate several features during each iteration. In *Group-wise Feature Selection* (GFS) [38], at each iteration, the feature group inducing the largest increase in the performance is selected and all variables inside it are added to the current subset, until a good performance is reached. The main limitation of all grouping methods is that an appropriate grouping mechanism needs to be chosen, which largely influences the final outcome.

2.2.3.2 Wrapper Methods

Wrapper methods employ different kind of learning algorithms to evaluate the feature subset.

Several FS procedures use C4.5 algorithm as classifier [40-43]. Particularly, the *Las Vegas Wrapper* (LVW), proposed by Liu and Setiono [40], is the wrapper version of the LVF depicted in the previous section. As for LVF, it starts from the entire set of features considered as the best subset and a random subset is extracted at each iteration. At each step, the current subset is compared with the best current one: if it has a number of features and a classification error lower than the current best, it is considered the new best subset. The procedure is iteratively repeated until the classification error does not change over a number of iterations chosen by user.

Rudy and Huan [44] developed a backward selection working with Artificial Neural Networks (ANNs). In this study, the variable producing the largest increase in the classification error at each iteration is deleted from the subset. Other application of ANNs used as supervised classifier to address FS can be found in literature [45-48].

A lot of wrapper methods adopt Support Vector Machine (SVM) in order to assess the feature subsets performances [47, 49-56] and some of them employ SVM in combination with Genetic Algorithms (GAs) [57]. Specifically, in [47] a classical GA procedure is hybridized with SVM or ANNs, depending on whether they are dealing with the two-class or the maximum depth problem.

Other approaches involve Radial Basis Functions [58], Nearest-Neighbor [59-62], linear classifiers [63, 64], AQ14 [65], Bayesian Networks [66] and, maximum likelihood [67].

2.2.3.3 Embedded Methods

Most of the methods belonging to this category are essentially based on SVM learning algorithm [68-75]. In [68] a backward strategy is employed in which, during the SVM training, the less important features are iteratively removed from the current subset. Moreover, two versions of the algorithm are proposed, that use a linear kernel and a non-linear kernel for SVM [69].

Few other embedded methods use kernel Fisher discriminant analysis [76], ANNs [77] and linear models [78].

2.2.4 Unsupervised FS

Although FS strategies were initially developed exclusively in supervised conditions, with the growing of datasets containing unlabeled samples also *unsupervised FS* techniques were proposed. In this case, as the class label is not provided, the performances of supervised learning algorithms cannot be used to evaluate the subset goodness.

Similarly to the supervised approach, in unsupervised contexts not every feature contains information useful for grouping similar objects into classes or *clusters*, but some of them can be redundant or irrelevant. Moreover, a too high number of irrelevant features can misguide the unsupervised learning process [10]. In fact, it is expected that the selected features subset will produce an increase in the performances of the unsupervised learning in terms of efficiency and quality.

As unlabeled samples are available in this situation, unsupervised problems result more difficult than supervised ones due to two aspects [79]:

1. *The number of clusters depends on the cardinality of the feature subset* and consequently it could change modifying the dimensions of the space in which the samples are projected. An example of this situation can be observed in Figure 2.6, in which in one-dimensional space only two clusters are identified (a), while in two dimensions it is possible to recognize three different clusters (b). From this consideration, it is clear that it could be not correct to fix the number of clusters but it is better to include the finding of this parameter in the search for the best feature subset with wrapper methods.
2. *Some evaluation criteria could be biased with respect to the dimensionality*. Dy and Brodley [79] proved that there are evaluation criteria which increase or decrease their value with the increase of the space dimensions (e.g. maximum likelihood and scatter separability). Hence, a criterion normalized with respect to the subset cardinality should be preferred.

Although the research in the supervised field is broader, also in the context of unsupervised FS many algorithms were proposed in literature. The aim of the following sections is to give an overview of the most relevant methods, grouped according to the search process they use. As no training data are available in unsupervised problems, no embedded procedures were proposed in literature. Indeed, they need a set of labeled samples to be used during the training phase of the learning algorithm.

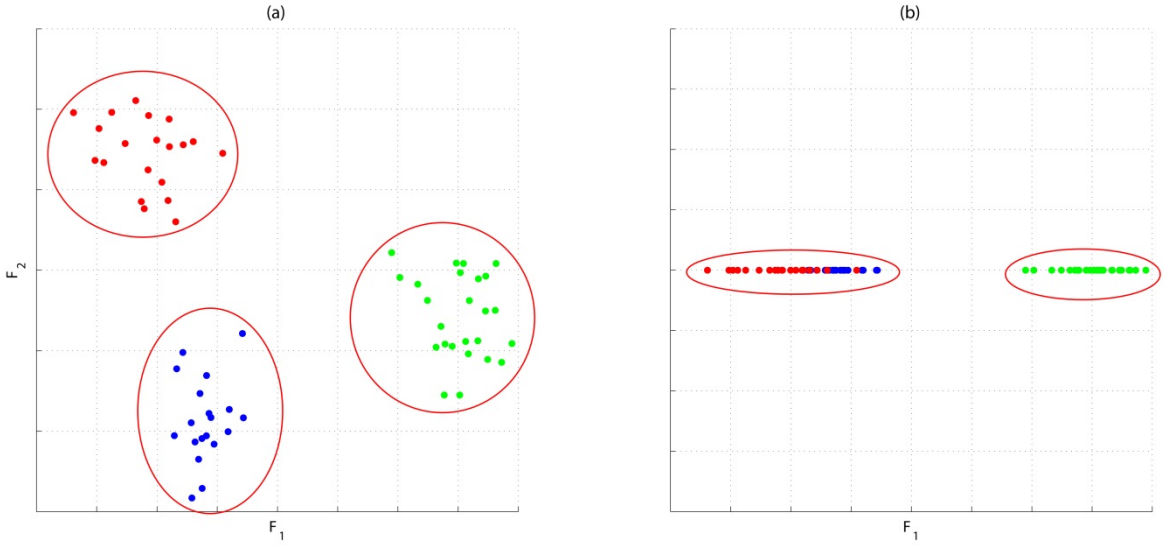

Figure 2.6 Illustration of how the number of clusters varies with dimensions. (a) Three different clusters are recognizable in two dimensions, (b) only two clusters can be observed in a two-dimensional space.

2.2.4.1 Filter Methods

Similarly to the supervised methods, different metrics were used for assessing the goodness of a feature subset, mainly following a heuristic search.

The FS algorithm proposed in [16] starts from the consideration that the entropy measure of a dataset is low if data has well-defined clusters, because the uncertainty is low, and high otherwise. Therefore, using a forward selection procedure several possible subsets of features are generated and evaluated by means of entropy. The research stops when the optimal subset is identified, that is the subset for which the entropy is minimum. A similar approach, based on forward selection and entropy evaluation can be found in [80]. The entropy measure is suitable for selecting the most important subset of features because it is invariant with number of dimensions, and is affected only by the quality of clustering.

Talavera [81] developed a FS method in which a weight is individually computed for each feature and only the k features with the highest weights are selected. As weight measure the function introduced by Fisher [82] is used, according to which a feature is relevant if knowing its value allows to improve the prediction of the value of other features. Other applications of dependency or correlation measures can be found in [83, 84].

One of the most widely employed distance metrics for unsupervised FS is the Laplacian Score [85]. It is based on the observation that, two data points

probably belong to the same cluster if they are close to each other. This metric, started by a k-means clustering method to select the top *k* features [85], is then applied to an iterative algorithm [86] and in combination with other measures such as entropy [87].

In *Unsupervised QuickReduct* algorithm, introduced in [88], a forward heuristic selection is performed using the Rough Set concept of *dependency degree* that is a measure of the consistency of the dataset, considering each feature with respect to all the others. The process iteratively adds to the current subset those producing a larger increase in the dependency degree, until the maximum value is reached.

2.2.4.2 Wrapper Methods

Several clustering methods were used in wrapper algorithms in order to find the feature subset able to highlight interesting natural groups in data. In this case, two objective functions can be used: the first one, based on the problem domain, for the assessment of the clustering and the second one for the evaluation of the feature subset according to those aspects the user is interested in [10].

In [89], the *Feature Subset Selection wrapped around EM clustering* (FSSEM)was formulated. The basic idea is to assess the goodness of each subset by firstly clustering data using expectation-maximization (EM) method [90] and then evaluating the resulting clusters and feature subset using two different criteria: Scatter Separability Criterion and Maximum Likelihood Criterion. Moreover, an improvement of the algorithm was proposed, *FSSEM-k* [89], in which for a given feature subset, also the number of clusters is identified.

Devaney and Ram [91] apply sequential forward and backward search, using COBWEB [82] as hierarchical clustering algorithm and evaluating each candidate subset by means of the category utility measure [92].The same clustering algorithm was used by Talavera [81] in a wrapper approach similar to the filter process described in 2.2.4.1.

A distributional clustering heuristic was adopted in [93], coupled with a probabilistic objective function based on Bayesian approach. The aim is to provide heuristics that find the optimum (or at least a sub-optimum) of this objective function in terms of the feature sets and the number of clusters.

Finally, in [94] an evolutionary local selection algorithm (ELSA)was formulated so that both the feature subset and the number of clusters are represented in each evolved solution. Two clustering algorithms, K-means and EM, are applied to form the given number of clusters based on the selected features.

Moreover, multiple fitness criteria are considered simultaneously for evaluating clustering models using a multi-objective or Pareto optimization.

References

1. Fayyad, U., G. Piatetsky-shapiro, and P. Smyth, *From Data Mining to Knowledge Discovery in Databases.* AI Magazine, 1996. **17**: p. 37-54.
2. Piatetsky-Shapiro, G., *Knowledge discovery in real databases: A report on the IJCAI-89 Workshop.* AI Magazine, 1991. **11**(5): p. 68-70.
3. Frawley, W.J., P. Shapiro, and C.J. Matheus, *Knowledge discovery in databases - an overview.* AI Magazine, 1992. **13**: p. 57-70.
4. Triantaphyllou, E., *Data Mining and Knowledge Discovery via Logic-Based Methods: Theory, Algorithms, and Applications*. 2010, New York, NY: Springer Publishing Company, Incorporated. 350.
5. Fayyad, U.M., G. Piatetsky-Shapiro, and P. Smyth, *From data mining to knowledge discovery: an overview*, in *Advances in knowledge discovery and data mining*, M.F. Usama, et al., Editors. 1996, American Association for Artificial Intelligence. p. 1-34.
6. Brachman, R.J. and T. Anand, *The process of knowledge discovery in databases*, in *Advances in knowledge discovery and data mining*, M.F. Usama, et al., Editors. 1996, American Association for Artificial Intelligence. p. 37-57.
7. Jensen, R. and Q. Shen, *Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches*. 2008, Hoboken, NJ: Wiley-IEEE Press. 340.
8. Cunningham, P., *Dimension Reduction*, in *Machine Learning Techniques for Multimedia*, M. Cord and P. Cunningham, Editors. 2008, Springer Berlin Heidelberg. p. 91-112.
9. Bellman, R., *Adaptive Control Processes: A Guided Tour*. 1961: Princeton University Press.
10. Jennifer, G.D., *Unsupervised Feature Selection*, in *Computational Methods of Feature Selection*, 2007, Chapman and Hall/CRC. p. 19-39.
11. Bolón-Canedo, V., N. Sánchez-Maroño, and A. Alonso-Betanzos, *A review of feature selection methods on synthetic data.* Knowledge and Information Systems, 2012: p. 1-37.
12. Jelonek, J. and J. Stefanowski, *Feature subset selection for classification of histological images.* Artificial Intelligence in Medicine, 1997. **9**(3): p. 227-239.
13. Singh, S., M. Singh, and M. Markou. *Feature selection for face recognition based on data partitioning*. in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*. 2002.
14. Yang, Y. and J.O. Pedersen, *A Comparative Study on Feature Selection in Text Categorization*, in *Proceedings of the Fourteenth International Conference on Machine Learning*1997, Morgan Kaufmann Publishers Inc. p. 412-420.

15. Forman, G., *An extensive empirical study of feature selection metrics for text classification.* J. Mach. Learn. Res., 2003. **3**: p. 1289-1305.
16. Dash, M., et al. *Feature selection for clustering - a filter solution*. in *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*. 2002.
17. Law, M.H.C., M.A.T. Figueiredo, and A.K. Jain, *Simultaneous feature selection and clustering using mixture models.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2004. **26**(9): p. 1154-1166.
18. Shang, C. and Q. Shen, *Aiding Classification of Gene Expression Data with Feature Selection: A Comparative Study.* International Journal of Computational Intelligence Research, 2005. **1**(1).
19. Saeys, Y., I. Inza, and P. Larranaga, *A review of feature selection techniques in bioinformatics.* Bioinformatics, 2007. **23**(19): p. 2507-17.
20. Au, W.H. and K.C.C. Chan. *An effective algorithm for discovering fuzzy rules in relational databases*. in *Fuzzy Systems Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on*. 1998.
21. Chen, S.-M., S.-H. Lee, and C.-H. Lee, *A new method for generating fuzzy rules from numerical data for handling classification problems.* Applied Artificial Intelligence, 2001. **15**(7): p. 645-664.
22. Yaochu, J., *Fuzzy modeling of high-dimensional systems: complexity reduction and interpretability improvement.* Fuzzy Systems, IEEE Transactions on, 2000. **8**(2): p. 212-221.
23. Shen, Q. and R. Jensen, *Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring.* Pattern Recognition, 2004. **37**(7): p. 1351-1363.
24. Somol, P., J. Novovi, and P. Pudil, *Notes on the Evolution of Feature Selection Methodology*. Kybernetika, 2007. **43**(5): p. 713-730.
25. Dash, M. and H. Liu, *Feature selection for classification.* Intelligent Data Analysis, 1997. **1**(1-4): p. 131-156.
26. Guyon, I., et al., *An introduction to variable and feature selection.* J. Mach. Learn. Res., 2003. **3**: p. 1157-1182.
27. Huan, L. and Y. Lei, *Toward integrating feature selection algorithms for classification and clustering.* IEEE Transactions on Knowledge and Data Engineering, 2005. **17**(4): p. 491-502.
28. Kira, K. and L. Rendell. *The Feature Selection Problem: Traditional Methods and a New Algorithm*. in *AAAI*. 1992. AAAI Press and MIT Press.
29. Kononenko, I., *Estimating attributes: analysis and extensions of RELIEF*, in *Proceedings of the European conference on machine learning on Machine Learning*1994, Springer-Verlag New York, Inc.: Catania, Italy. p. 171-182.
30. Liu, H., H. Motoda, and L. Yu, *Feature Selection with Selective Sampling*, in *Proceedings of the Nineteenth International Conference on Machine Learning*2002, Morgan Kaufmann Publishers Inc. p. 395-402.
31. Segen, J. *Feature selection and constructive inference*. 1984. Montreal, Que, Can: IEEE.

32. Hanchuan, P., L. Fuhui, and C. Ding, *Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2005. **27**(8): p. 1226-1238.
33. Hall, M., *Correlation-based Feature Subset Selection for Machine Learning*, 1999, Department of Computer Science, University of Waikato.
34. Seth, S. and J.C. Principe. *Variable Selection: A Statistical Dependence Perspective*. in *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*. 2010.
35. Almuallim, H. and T.G. Dietterich. *Learning with many irrelevant features*. in *Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI-91)*. 1991. AAAI Press.
36. Liu, H. and R. Setiono, *A Probabilistic Approach to Feature Selection - A Filter Solution*, in *ICML'96*1996. p. 319-327.
37. Jensen, R. and S. Qiang, *Using fuzzy dependency-guided attribute grouping in feature selection*, in *Proceedings of the 9th international conference on Rough sets, fuzzy sets, data mining, and granular computing*2003, Springer-Verlag: Chongqing, China. p. 250-254.
38. Pacl, P., et al., *On Feature Selection with Measurement Cost and Grouped Features*, in *Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*2002, Springer-Verlag. p. 461-469.
39. Bondell, H.D. and B.J. Reich, *Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR.* Biometrics, 2008. **64**(1): p. 115-123.
40. Liu, H. and R. Setiono, *Feature Selection and Classification - A Probabilistic Wrapper Approach*, in *IEA/AIE*, T. Tanaka, S. Ohsuga, and M. Ali, Editors. 1996, Gordon and Breach Science Publishers. p. 419-424.
41. John, G., R. Kohavi, and K. Pfleger. *Irrelevant Features and the Subset Selection Problem*. in *International Conference on Machine Learning*. 1994.
42. Caruana, R. and D. Freitag, *Greedy Attribute Selection*, in *Proceedings of the Eleventh International Conference on Machine Learning*1994, Morgan Kaufmann. p. 28-36.
43. Chen, Y., L. Dai, and X.Q. Cheng. *GATS-C4.5: An algorithm for optimizing features in flow classification*. 2008. Las Vegas, NV.
44. Rudy, S. and L. Huan, *Neural-network feature selector.* Neural Networks, IEEE Transactions on, 1997. **8**(3): p. 654-662.
45. Verikas, A. and M. Bacauskiene, *Feature selection with neural networks.* Pattern Recognition Letters, 2002. **23**(11): p. 1323-1335.
46. Boehm, O., D.R. Hardoon, and L.M. Manevitz, *Towards one-class pattern recognition in brain activity via neural networks*, 2010: Pachuca. p. 126-137.
47. Villar, J., et al., *Soft Computing Decision Support for a Steel Sheet Incremental Cold Shaping Process*, in *Intelligent Data Engineering and Automated Learning - IDEAL 2011*, H. Yin, W. Wang, and V. Rayward-Smith, Editors. 2011, Springer Berlin / Heidelberg. p. 482-489.

48. Yang, J. and V.G. Honavar, *Feature Subset Selection Using a Genetic Algorithm.* IEEE Intelligent Systems, 1998. **13**(2): p. 44-49.
49. Pan, H., L.Z. Xia, and T.Q. Nguyen. *Robust object detection scheme using feature selection*. in *Proceedings of 2010 IEEE 17th International Conference on Image Processing*. 2010. Hong Kong.
50. Kramer, K., et al. *Increased classification accuracy and speedup through pair-wise feature selection for support vector machines*. in *Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on*. 2011.
51. Song, X., et al., *Feature selection for support vector machine in financial crisis prediction: a case study in China.* Expert Systems, 2010. **27**(4): p. 299-310.
52. Durbha, S.S., R.L. King, and N.H. Younan, *Wrapper-Based Feature Subset Selection for Rapid Image Information Mining*. Geoscience and Remote Sensing Letters, IEEE, 2010. **7**(1): p. 43-47.
53. Aggarwal, N., R.K. Agrawal, and H.M. Jain. *Genetic algorithm to determine relevant features for intrusion detection*. in *IADIS European Conference Data Mining 2009*. 2009. Algarve.
54. Maldonado, S. and R. Weber, *A wrapper method for feature selection using Support Vector Machines.* Information Sciences, 2009. **179**(13): p. 2208-2217.
55. Zhuo, L., et al. *A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine*. in *Proceedings of SPIE - The International Society for Optical Engineering*. 2008. Guangzhou.
56. Alba, E., et al. *Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms*. in *Evolutionary Computation, 2007. CEC 2007. IEEE Congress on*. 2007.
57. Holland, J.H., *Adaptation in natural and artificial systems*. 1992: MIT Press. 211.
58. Xiuju, F. and W. Lipo. *A GA-based RBF classifier with class-dependent features*. in *Evolutionary Computation, 2002. CEC '02. Proceedings of the 2002 Congress on*. 2002.
59. Brill, F.Z., D.E. Brown, and W.N. Martin, *Fast generic selection of features for neural network classifiers.* Neural Networks, IEEE Transactions on, 1992. **3**(2): p. 324-328.
60. Aha, D.W. and R.L. Bankert. *Feature selection for case-based classification of cloud types: An empirical comparison*. in *Proceedings of the 1994 {AAAI} Workshop on Case-Based Reasoning*. 1994. AAAI Press.
61. Langley, P. and S. Sage. *Oblivious decision trees and abstract cases*. in *Case-Based Reasoning: Papers from the 1994 Workshop (Technical Report WS-94-01)*. 1994. AAAI Press.
62. Li, L., et al., *Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method.* Bioinformatics, 2001. **17**(12): p. 1131-1142.
63. Foroutan, I. and J. Sklansky, *Feature selection for automatic classification of non-Gaussian data.* IEEE Trans. Syst. Man Cybern., 1987. **17**(2): p. 187-198.

64. Malarvili, M.B., M. Mesbah, and B. Boashash. *HRV Feature Selection for Neonatal Seizure Detection: A Wrapper Approach*. in *Signal Processing and Communications, 2007. ICSPC 2007. IEEE International Conference on*. 2007.
65. Vafaie, H. and I. Imam, *Feature Selection Methods: Genetic Algorithms vs. Greedy-like Search*, 1994.
66. Inza, I., et al., *Feature Subset Selection by Bayesian network-based optimization.* Artificial Intelligence, 2000. **123**(1–2): p. 157-184.
67. Ooi, C.H. and P. Tan, *Genetic algorithms applied to multi-class prediction for the analysis of gene expression data.* Bioinformatics, 2003. **19**(1): p. 37-44.
68. Guyon, I., et al., *Gene Selection for Cancer Classification using Support Vector Machines.* Mach. Learn., 2002. **46**(1-3): p. 389-422.
69. Rakotomamonjy, A., *Variable selection using svm based criteria.* J. Mach. Learn. Res., 2003. **3**: p. 1357-1370.
70. Neumann, J., C. Schnrr, and G. Steidl, *Combined SVM-Based Feature Selection and Classification.* Mach. Learn., 2005. **61**(1-3): p. 129-150.
71. Maldonado, S., R. Weber, and J. Basak, *Simultaneous feature selection and classification using kernel-penalized support vector machines.* Information Sciences, 2011. **181**(1): p. 115-128.
72. Maldonado, S. and R. Weber, *Embedded Feature Selection for Support Vector Machines: State-of-the-Art and Future Challenges*, in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, C. San Martin and S.-W. Kim, Editors. 2011, Springer Berlin / Heidelberg. p. 304-311.
73. Xu, Y., P. Zhong, and L. Wang, *Support vector machine-based embedded approach feature selection algorithm.* Journal of Information and Computational Science, 2010. **7**(5): p. 1155-1163.
74. Maldonado, S. and G. L'Huillier. *Embedded feature selection for spam and phishing filtering using support vector machines*. 2012. Vilamoura, Algarve.
75. Hernandez, J.C.H., B. Duval, and J.-K. Hao, *A genetic embedded approach for gene selection and classification of microarray data*, in *Proceedings of the 5th European conference on Evolutionary computation, machine learning and data mining in bioinformatics*2007, Springer-Verlag: Valencia, Spain. p. 90-101.
76. Wang, Y.Q. *Kernel fisher discriminant analysis embedded with feature selection*. 2007. Hong Kong.
77. Bo, L., L. Wang, and L. Jiao, *Multi-layer perceptrons with embedded feature selection with application in cancer classification.* Chinese Journal of Electronics, 2006. **15**(4 A): p. 832-835.
78. Keleş, S., M. van der Laan, and M.B. Eisen, *Identification of regulatory elements using a feature selection method.* Bioinformatics, 2002. **18**(9): p. 1167-1175.
79. Dy, J.G. and C.E. Brodley, *Feature Selection for Unsupervised Learning.* J. Mach. Learn. Res., 2004. **5**: p. 845-889.
80. Dash, M., H. Liu, and J. Yao. *Dimensionality reduction of unsupervised data*. in *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on*. 1997.

81. Talavera, L., *Feature Selection as a Preprocessing Step for Hierarchical Clustering*, in *Proceedings of the Sixteenth International Conference on Machine Learning*1999, Morgan Kaufmann Publishers Inc. p. 389-397.
82. Fisher, D.H., *Knowledge Acquisition Via Incremental Conceptual Clustering.* Mach. Learn., 1987. **2**(2): p. 139-172.
83. Mitra, P., C.A. Murthy, and S.K. Pal, *Unsupervised feature selection using feature similarity.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2002. **24**(3): p. 301-312.
84. Das, S.K., *Feature Selection with a Linear Dependence Measure.* Computers, IEEE Transactions on, 1971. **C-20**(9): p. 1106-1109.
85. He, X., D. Cai, and P. Niyogi, *Laplacian Score for Feature Selection*, in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Editors. 2006, MIT Press. p. 507-514.
86. Zhu, L., L. Miao, and D. Zhang, *Iterative Laplacian Score for Feature Selection*, in *Pattern Recognition*, C.-L. Liu, C. Zhang, and L. Wang, Editors. 2012, Springer Berlin Heidelberg. p. 80-87.
87. Rongye, L., et al. *An Unsupervised Feature Selection Algorithm: Laplacian Score Combined with Distance-Based Entropy Measure*. in *Intelligent Information Technology Application, 2009. IITA 2009. Third International Symposium on*. 2009.
88. Velayutham, C. and K. Thangavel, *Unsupervised Quick Reduct Algorithm Using Rough Set Theory.* Journal of Electronic Science and Technology, 2011. **9**(3): p. 193-201.
89. Dy, J.G. and C.E. Brodley, *Feature Subset Selection and Order Identification for Unsupervised Learning*, in *International Conference on Machine Learning*2000. p. 247-254.
90. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society. Series B (Methodological), 1977. **39**(1): p. 1-38.
91. Devaney, M. and A. Ram, *Efficient Feature Selection in Conceptual Clustering*, in *Proceedings of the Fourteenth International Conference on Machine Learning*1997, Morgan Kaufmann Publishers Inc. p. 92-97.
92. Gluck, M. and J. Corter. *Information, uncertainty, and the utility of categories*. in *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*. 1985. Lawrence Earlbaum.
93. Vaithyanathan, S. and B. Dom. *Model Selection in Unsupervised Learning With Applications To Document Clustering*. in *Proc. 16th International Conference on Machine Learning (ICML'99)*. 1999. Morgan Kaufmann.
94. Kim, Y., W.N. Street, and F. Menczer, *Evolutionary model selection in unsupervised learning.* Intell. Data Anal., 2002. **6**(6): p. 531-556.

Chapter 3: Materials & Methods

3.1 Computational Intelligence (CI)

Nowadays there is a growing need of developing tools and algorithms able to model and solve increasingly complex problems. In this sense, *Artificial Intelligence (*AI) provides a set of intelligent systems and techniques able to acquire knowledge, learn and act similarly to what humans daily do.

Computational Intelligence (CI) is a branch of AI dealing with intelligent tools and models inspired by biological and natural systems and employed to solve complex problems. In particular, CI was defined as "the study of adaptive mechanisms to enable or facilitate intelligent behavior in complex, uncertain and changing environment" [1, 2]. Thanks to their ability to learn and adapt to new situations, to generalize, abstract, discover and associate, CI techniques are successfully applied in several fields including biology, chemistry, computer science, economics, electromagnetism, engineering, immunology, information science, linguistics, material science, music and physics [1].

Three main classes of problems are faced by CI: optimization, supervised learning and unsupervised learning.

Optimization is undoubtedly the primary category of problems which CI research focuses on. Theoretically, any other class of problems can be reformulated as an optimization problem. The goal of optimization is to find the solution of a specific problem that maximizes or minimizes (generally that optimizes) a given objective function. The search of the best solution is usually led by a heuristic strategy allowing to reduce the search domain and to avoid the analysis of the complete set of possible solutions. Moreover, some constraints

can be introduced in the problem domain, that further reduce the feasible region of solutions.

Supervised learning is the process of learning the intrinsic properties of a system in terms of relationship between inputs and outputs, using a set of examples characterized by input variables and target output. Once trained, the learned system can provide the output of a previously unused example. Supervised learning can be used for different kind of problems such as classification, pattern recognition and control problems.

The last class of problems treated by means of CI tools is *unsupervised learning*, generally related to clustering problems. The aim is to find homogeneous groups of elements usually basing on a distance metric that gives a measure of the similarity or dissimilarity between two objects.

Initially, CI was defined as the combination of fuzzy logic, neural networks, and genetic algorithms [1]. Now, five paradigms are included into CI and are showed in Figure 3.1: Artificial Neural Networks (ANNs), Evolutionary Computation (EC), Swarm Intelligence (SI), Fuzzy Systems (FSys), and Artificial Immune Systems (AISs) [2]. These branches can be also combined among them in order to obtain hybrid algorithms, as shown in Figure 3.1.

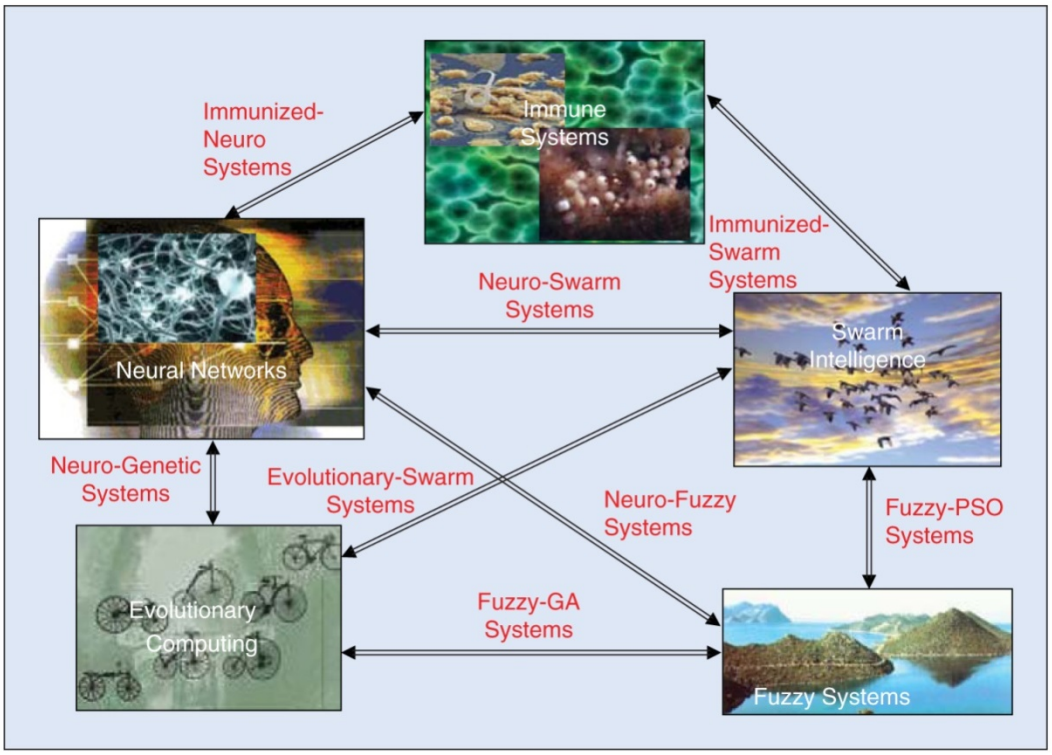

Figure 3.1 The five main paradigms of CI (Artificial Neural Networks, Evolutionary Computation, Swarm Intelligence, Fuzzy Systems, and Artificial Immune Systems) and their combinations to obtain hybrid algorithms [1].

The final purpose of all CI techniques is to provide solutions for real-world problems that could be human-interpretable, low cost, and exact-enough. The research in this field is going in the direction of developing more intelligent and robust systems for pattern classification, learning, optimization and decision-making.

Actually, there is a large interest in the application of intelligent methods to the biomedical field [3-5], aiming to provide physicians with Computer-Aided Diagnosis tools able to assist and support their decisions.

3.1.1 Artificial Neural Networks (ANN)

ANN is a mathematical model which simulates the biophysical information processing occurring in the nervous system. Similarly to the human brain, an ANN is able to learn, store and generalize knowledge from data. These characteristics make ANNs suitable for classification and clustering tasks, fitting, forecasting, optimization problems and information retrieval by contents.

The basic element of an ANN is the *Artificial Neuron* (AN) or *perceptron*, modeling the characteristics of a biological neuron. Miming what happens in the brain, each AN receives inputs from the environment or other neurons by means of weighted connections. The learning task is achieved through a learning algorithm that acts modifying the weights of the connections between neurons.

Essentially two categories of ANNs configurations are possible, according to the direction of the connections between the ANs: Feed-Forward Neural Networks (FFNN) (§3.6.1), in which each layer is only connected with the following one, and Feed-Back Neural Networks, in which information is propagated in both directions.

For classification purposes, usually more than a single perceptron is needed, producing ANNs made of sequences of neuron layers connected among them. Each AN output is obtained as a weighted combination of the inputs that is further elaborated by a nonlinear activation function specific for each neuron (Figure 3.2). The perceptron outcome is then transferred to all connected ANs until it reaches the output layer. Here, the error between the current output and the desired output is calculated and propagated backward to adjust the weights of the neuron connections.

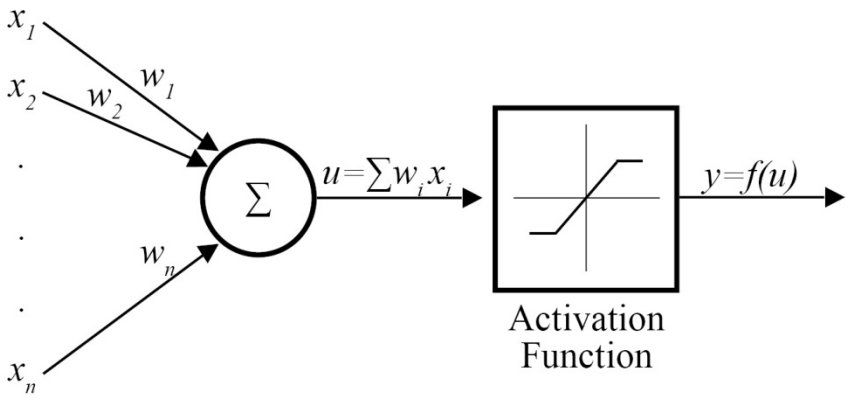

Figure 3.2 Example of an *Artificial Neuron* (AN) or *perceptron*.

For clustering tasks, the most used network typology is the *Kohonen's Self-Organized Map* (SOM). It is made of one input layer connected with a 2-D layer of ANs. In this case, the learning phase is based on the concept of *competition*: when a new training input is received by the SOM, neurons compete among them for activation and a winner neuron is selected, that is the one whose weights are the most similar to the input vector. Moreover, during the competitive learning both the weights associated to the winner neuron and the ones of the neighbor neurons (*neighborhood*) are updated. A representation of a SOM is showed in Figure 3.3.

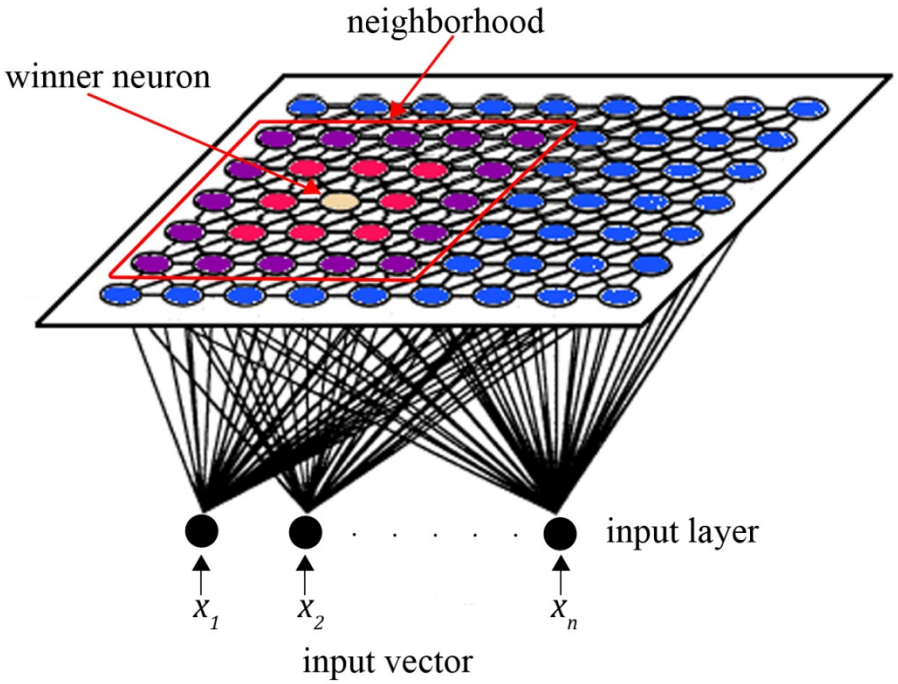

Figure 3.3 Representation of a *Kohonen's Self Organized Map* (SOM).

A particular case of Multi-Layer Perceptron is represented by the *Radial Basis Function Neural Network* (RBFNN), combining supervised and unsupervised learning. A RBFNN is a FFNN made of one input layer, one hidden layer and an output layer with linear activation function. Each hidden unit implements a radial basis function depending on the distance between the input and the center of the function. These centers are estimated, during the training process, by means of an unsupervised learning algorithm (such as the k-means). Once all centers are identified, the neuron weights are finally updated, according to the distance between the actual and the desired output.

3.1.2 Evolutionary Computation (EC)

EC includes a set of tools inspired by the Darwin theory about the natural evolution of species. The main idea is that, as it happens in natural selection, only the best characteristics for adapting to the surrounding environment are transmitted to the new generations. This is obtained through the selection of the best individuals, which will be able to reproduce and transmit their winning genes.

In EC, each individual (*chromosome*) represents a possible solution of the problem. A *fitness function* is associated to each individual and represents its goodness in solving the problem. During the algorithm, the population evolves so that the best individuals (the ones with the highest value of fitness) have a high probability to reproduce in order to generate a new population of offspring. The common evolutionary operators, *crossover* and *selection*, effect the population evolution, until a stopping criterion is satisfied (generally a fixed number of iterations or an acceptable fitness value).

The most important application of this class of CI tools is related to optimization problems. However, it is also used for classification, clustering, feature selection and related applications.

Different modifications to the original EC algorithms were proposed, producing several classes of tools such as Genetic Algorithms (GAs), Genetic Programming, Evolutionary Programming, Evolutionary Strategies, Differential Evolution, Cultural Evolution, and Coevolution [2].

3.1.3 Swarm Intelligence (SI)

SI algorithms are inspired by the social behavior of those species living in swarm (ants, bees, …) and model the social interactions among individuals while performing specific tasks, such as searching food. Similarly to the EC

strategies, this class of tools is based on a population of organisms representing the candidate solutions to the problem, and uses a fitness function, formulated according to the specific application, to measure the performance of each individual. Optimization and clustering problems are commonly solved applying SI tools.

The most important algorithms belonging to the SI category are essentially two: Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO).

PSO draws on the behavior of birds living in flocks. Each possible solution (called *particle*) moves in the search space according to its own experience and to the knowledge acquired by other particles.

ACO mimics the same strategy used by ants in order to find the best path in the direction of food. Each route followed by ants represents a feasible solution to the real problem. The main idea is that each ant, during its way from the colony nest to the food source, deposits on the ground a chemical hormone called *pheromone*. This substance helps other ants in the choice of the best trail to reach the food and consequently to find the best solution to the problem: the more ants pass on a path, the stronger is the pheromone trace, which identify the path as the best one.

3.1.4 Fuzzy Systems (FSys)

In the traditional crisp set theory an element can either belong to a set or not. This means that the membership degree of an item to a specific set is defined as a Boolean value in which 0 represents an object that is not included in the set, and 1 indicates that the object belongs to the set. Usually human reasoning is not so exact in classifying an element, but it includes a certain degree of uncertainty that is not representable using the traditional set theory.

FSys allow modeling all those situations in which an object belongs to a class with an associated degree of membership different from one. Consequently, using FSys it is possible to reproduce the common reasoning including vagueness, imprecision and ambiguity. These abilities enable also to describe rules governing a given systems from a fuzzy point of view, using domain specific terms and natural language, which can be better explained and understood.

FSys includes Fuzzy Logic, Fuzzy Inference Systems (§3.2), Fuzzy Sets and Rough Sets (§3.3).

3.1.5 Artificial Immune Systems (AISs)

AISs are rather new CI approaches inspired by the ability of the lymphocytes in the human immune system to distinguish between pathogens entering in the body and human cells. Once lymphocytes detect a specific antigen, they will be able to recognize and respond to it faster in case of future events.

In AISs, instead of obtaining a model of the training data (that are representative of the system to be modelled), an implicit classifier is built, modeling everything else but the training data. In this way the AIS is able to detect anomalous behaviors in systems.

Because of their intrinsic characteristics, AISs are mainly applied to pattern recognition and classification problems, in particular for anomaly detection such as fraud detection and computer virus detection.

3.2 Fuzzy Inference Systems (FIS)

A Fuzzy Inference System (FIS) is a method for mapping from a given set of inputs to a certain number of outputs, based on Fuzzy Theory concepts. It aims to mime the same procedures of the human reasoning, which is characterized by a certain degree of vagueness and uncertainty, by means of *fuzzy sets* and *if-then rules*.

3.2.1 Fuzzy Sets

A *fuzzy set A* is a set with fuzzy boundaries, defined by a function $\mu_A(x): x \rightarrow [0,1]$ called *membership function* (MF). It associates to the element x a certain degree of membership to the fuzzy set A that can range from 0 to 1. If x completely belongs to the set A then$\mu_A(x) = 1$, if x is not in A then $\mu_A(x) = 0$, if x is partially included in A then $0 < \mu_A(x) < 1$. This kind of reasoning is impossible to be obtained with the classical crisp logic, which allows only two values of membership to a set, 0 and 1.

Fuzzy sets are associated to linguistic variables used for describing real world concepts. For example, the variable *height*, describing the height of a man, can be modeled in fuzzy-terms using three sets: *short, normal*, and *tall*. Each set is defined by a MF, which associates to a certain input x representing the man height a given membership degree to the fuzzy sets, as showed in Figure 3.4. In this example, a man with height 184 cm belongs to the set of *tall* with membership degree of 0.4 and, in the same time, to the set of *normal* with a degree of 0.1.

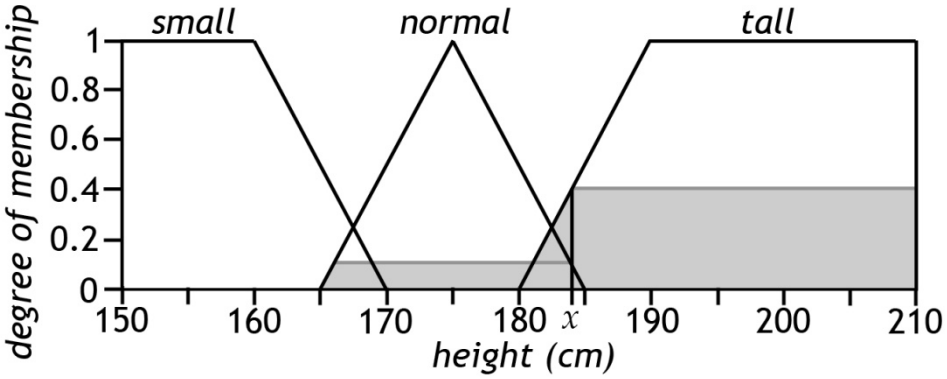

Figure 3.4 Membership functions associated to the fuzzy sets *small, normal* and *tall* describing the variable *height*. In the example, a man with height 184 cm (x) belongs to the set of *tall* with membership degree of 0.4 and, in the same time, to the set of *normal* with a degree of 0.1.

3.2.2 Fuzzy Rules

Fuzzy rules are linguistic rules expressing knowledge, experience or expertise about the domain of analysis. They link a set of *antecedents* with some *consequents,* in the general form of:

$$if\ x\ is\ A\ (antecedent) then\ y\ is\ B\ (consequent)$$

where x and y are linguistic variables and A and B are the fuzzy sets associated to the universe of discourse of x and y, respectively.

In the classical rule-based logic, if the antecedent is true, then the consequent is also true. In fuzzy logic, if the antecedent is true with a certain degree of membership, then the consequent is also true with the same degree. For example, in the case of *men's height* a fuzzy rule can express the relation between height and weight as:

$$if\ height\ is\ tall\ then\ weight\ is\ heavy.$$

Using the example of a man 184 cm tall, as he belonged to the *tall* set with a membership degree of 0.4, he belongs also to the *heavy*set with the same degree of membership 0.4.

Moreover, fuzzy rules can admit multiple antecedents and consequents, that are combined through the application of fuzzy operators such as complement (NOT), intersection (AND) or union (OR).

The complement of a fuzzy set A is a set $\bar{A}$ containing all elements of A with a membership degree that is $\mu_{\bar{A}} = 1 - \mu_A$.

The intersection of two fuzzy sets A and B is the set of elements contained in both sets, and it is usually implemented using the *min-operator* as: $\mu_{A\cap B} = \min\{\mu_A, \mu_B\}$.

Finally, the union of two fuzzy sets A and B is the set containing all elements of A and B, and it is most frequently obtained applying the *max-operator* as: $\mu_{A\cup B} = \max\{\mu_A, \mu_B\}$.

3.2.3 Fuzzy Inference

A FIS is a system made of a given number of input and output variables, described by a certain number of fuzzy sets and connected through a list of fuzzy rules.

Two inference approaches were proposed in literature, the Mamdani inference and the Sugeno inference [6]. The most widely used is the first one, formulated in 1975 by Professor Mamdani of London University [7].

The Mamdani inference process can be mainly divided into four steps:

1. *Fuzzification*: In this step, the crisp inputs are transformed in fuzzy-representations, by determining their membership degree to each appropriated fuzzy set of inputs.
2. *Rule Evaluation*: starting from the fuzzified inputs, the antecedents of each rule are evaluated and the results are applied to the consequents. This means that the output set is activated with the same degree of membership than the antecedent. If multiple antecedents are included in the same rules, the fuzzy operators (AND or OR) are used to obtain the membership degree of the output. Moreover, fuzzy rules can be weighted between 0 and 1 to represent the degree of confidence in the rule. These weights are determined a priori by the expert during the design of the FIS. This step leads to obtain a series of fuzzified outputs, one for each rule.
3. *Aggregation*: this phase allows combining the fuzzified outputs of all rules to obtain a single fuzzy set. Fuzzy operators are used to aggregate the outputs, similarly to the process for the antecedents' combination.
4. *Defuzzification*: to obtain a crisp number as FIS output, several defuzzification methods can be applied to the aggregate fuzzy output. The most popular is the centroid or the center of gravity, which returns as crisp output the horizontal coordinate of the center of gravity of the aggregated set.

The basic structure of the Mamdani inference process is depicted in Figure 3.5.

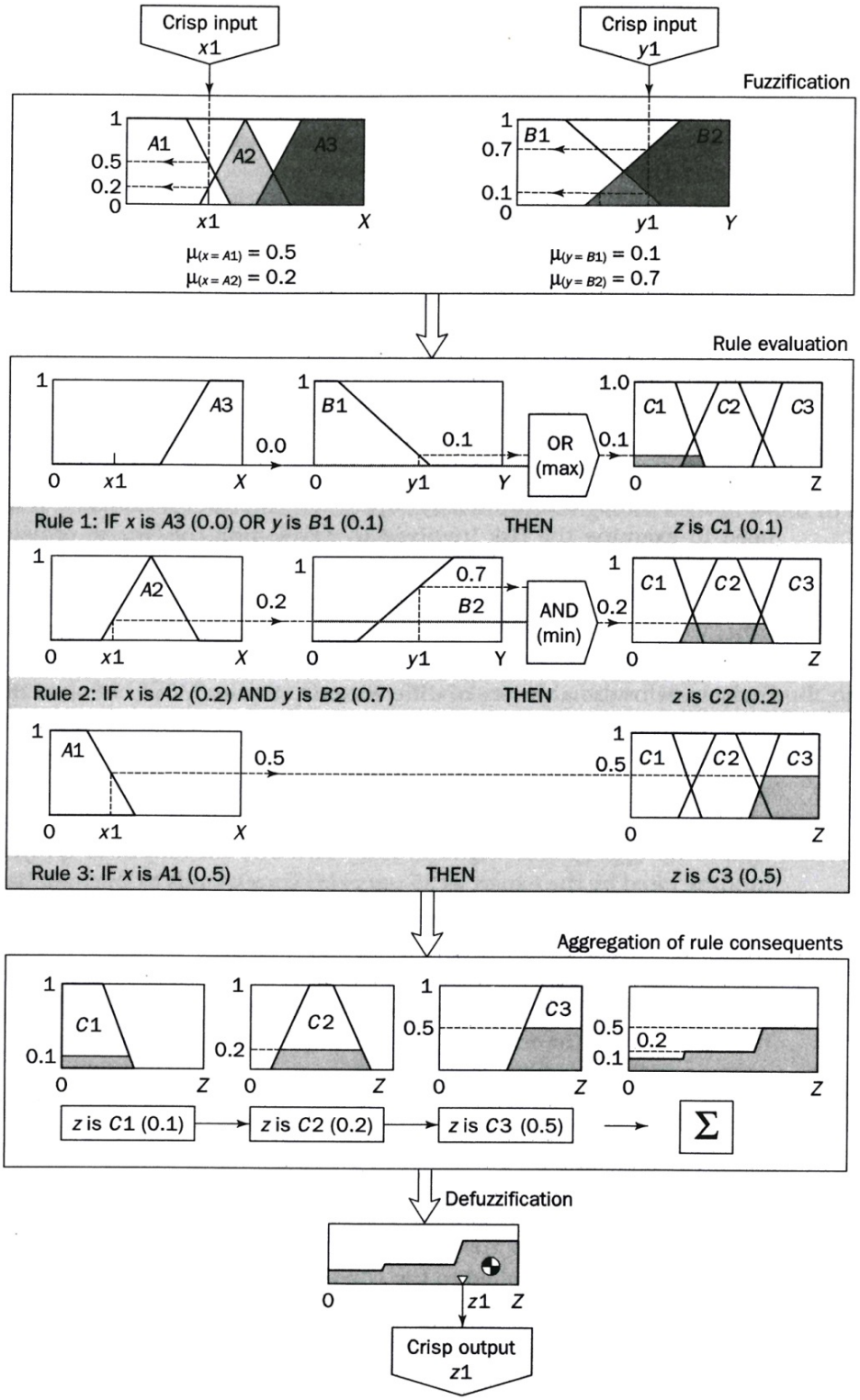

Figure 3.5 Basic structure of the Mamdani inference process [6].

3.3 Rough Set Theory (RST)

Rough Set Theory (RST), introduced by Pawlakin the 1980s [8], is a powerful tool for managing uncertainty afflicting the clinical domain in efficient and effective manner. Unlike the classical CI techniques, RST does not require any a-priori information or model assumption about data, but it uses only knowledge directly derivable from the given data. Moreover, RST is able to model imperfect and incomplete knowledge, which usually characterizes medical datasets.

RST was born as extension of fuzzy theory. Although both approaches allow managing uncertainty, the type of uncertainty they deal with is different: Fuzzy Theory is related to vagueness while RST concerns with indiscernibility.

In the past years RST has found wide and different areas of application, such as classification [9], system monitoring [10], clustering [11], expert systems [12]and decision analysis [13], but the field in which it significantly contributed is KDD. In this research domain, RST was successfully employed for discovering data dependencies, evaluating features importance, discovering patterns in data, removing redundant objects and features, recognizing and classifying elements and extracting readable if-then rules[14].

In particular, FS was one of the most important fields in which RST has been applied [15-17], with very satisfactory results.

3.3.1 Foundations

In RST, data are organized in a *decision system* or *decision table:* the rows are the objects in the *universe* $\mathbb{U}$ and columns represent the *attributes* $\mathbb{A}$ characterizing each element. Two kinds of attributes are identified: *conditional attributes* $\mathbb{C}$, which represent the features or characteristics of each object, and a *decision attribute* $\mathbb{D}$, which is the class the objects belong to. An example of a decision system is reported in Table 3.I, where there are eight objects x described by four conditional features (a,b,c,d) and one decision feature (e).

The basic principle of RST is *indiscernibility*: if two objects are indiscernible with respect to a certain feature, then they should be classified in the same class. Mathematically, this concept can be written for a non-empty subset of features $P \subseteq \mathbb{A}$ as:

$$IND(P) = \{(x, y) \in \mathbb{U}^2 : \forall\, a \in P, a(x) = a(y)\} \quad (3.1)$$

where $a \in P$ denotes the value of the attributes in P for a specific object and (x, y) is a pair of objects indiscernible with respect to the P attributes. $IND(P)$ determines a partition of the universe $\mathbb{U}$, identified as $\mathbb{U}/IND(P)$, which contains

the set of *equivalence classes* $[x]_P$ with respect to P. For the example in Table 3.I, the partition of the universe created by $P = \{b, c\}$ is made of the subsets of objects indiscernible with respect to *b* and *c*, and is the following:

$$\mathbb{U}/IND(P) = \bigcup [x_P] = \{\{3\}, \{1,5\}, \{4\}, \{2,7,8\}, \{6\}\}.$$

Table 3.I An example of a *decision table* or *decision system* containing eight objects *x* characterized by four *conditional features* (*a,b,c,d*) and a *decision feature* (*e*).

Object x	**a**	**b**	**c**	**d**	**e**
1	S	R	T	T	R
2	R	S	S	S	T
3	T	R	R	S	S
4	S	S	R	T	T
5	S	R	T	R	S
6	T	T	R	S	S
7	T	S	S	S	T
8	R	S	S	R	S

According to RST, each subset of objects $X \subseteq \mathbb{U}$ can be divided into two disjoint regions, named *lower* and *upper approximations*, using only the information contained in P and defined respectively as:

$$\underline{P}X = \{x \in \mathbb{U} | [x]_P \subseteq X\}$$

$$\overline{P}X = \{x \in \mathbb{U} | [x]_P \cap X \neq \emptyset\}.$$

The *lower approximation* $\underline{P}X$ contains all equivalence classes $[x]_P$ that are certainly subsets of X, while the *upper approximation* $\overline{P}X$ includes all equivalence classes $[x]_P$ having objects in common with X. A graphic representation of these two concepts is reported in Figure 3.6.

Starting from the two approximation concepts, three different regions can be identified, that are graphically illustrated in Figure 3.6.

The *positive region* $POS_P(D)$ includes all the objects of universe$\mathbb{U}$ that can be uniquely classified into $\mathbb{U}/IND(D)$ by using only the attributes of *P*:

$$POS_P(D) = \bigcup_{X \in \mathbb{U}/IND(D)} \underline{P}X.$$

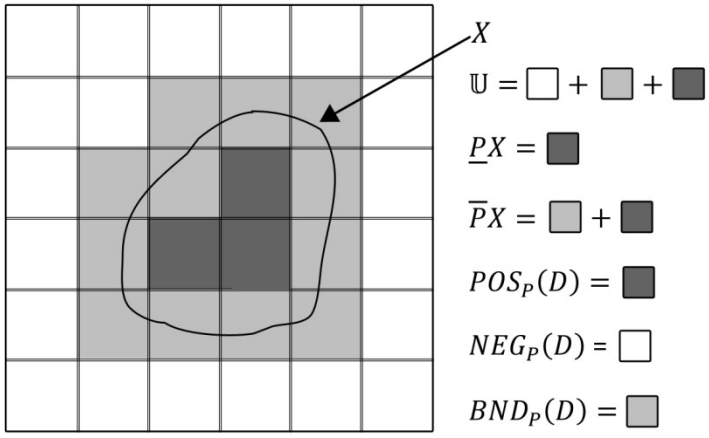

Figure 3.6 Graphic representation of the main *Rough Set Theory* (RST) concepts [18].

The *negative region* $NEG_P(D)$ is the complete set of objects that certainly does not belong to the $\mathbb{U}/IND(D)$ classes, according to the attributes in P:

$$NEG_P(D) = \mathbb{U} - \bigcup_{X \in \mathbb{U}/IND(D)} \bar{P}X.$$

Finally, the *boundary region* $BND_P(D)$ contains objects that can possibly, but not certainly, belong to X:

$$BND_P(D) = \bigcup_{X \in \mathbb{U}/IND(D)} \bar{P}X - \bigcup_{X \in \mathbb{U}/IND(D)} \underline{P}X.$$

Referring to the example in Table 3.I and considering $P = \{b, c\}$ and $D = \{e\}$ the three regions are defined as follows:

$$POS_P(D) = \cup\{\emptyset, \{3,6\}, \{4\}\} = \{3,4,6\}$$

$$NEG_P(D) = \mathbb{U} - \cup\{\{1,5\}, \{3,1,5,2,7,8,6\}, \{4,2,7,8\}\} = \emptyset$$

$$BND_P(D) = \cup\{\{1,5\}, \{3,1,5,2,7,8,6\}, \{4,2,7,8\}\} - \cup\{\emptyset, \{3,6\}, \{4\}\} = \{1,2,5,7,8\}$$

Let $P \subseteq \mathbb{C}$ be a subset of conditional features and $D \subseteq \mathbb{D}$ the decision feature, it is possible to measure the importance of P in classifying the objects x into the class D by means of the *dependency degree*:

$$\gamma_P(D) = \frac{|POS_P(D)|}{|\mathbb{U}|}$$

where $|\cdot|$ denotes the cardinality of set. If $\gamma_P(D) = 1$, the classification D totally depends on the attributes in P, whereas a values of $\gamma_P(D)$ between zero and one means that D partially depends on P. If $\gamma_P(D)$ is equal to 1, the dataset is de-

fined as *consistent*, i.e. for every set of objects whose conditional features values are the same, the corresponding decision attributes are identical. The *dependency degree* for the dataset in Table 3.I, considering $P = \{b, c\}$ and $D = \{e\}$, is given by:

$$\gamma_P(D) = \frac{|\{3,4,6\}|}{|\{1,2,3,4,5,6,7,8\}|} = \frac{3}{8}$$

Moreover, it is possible to calculate the *significance* $\sigma_P(D)$ of a conditional feature by measuring the change in the dependency degree when the selected feature is removed from the subset of conditional features:

$$\sigma_P(D) = \gamma_P(D) - \gamma_{P-\{a\}}(D).$$

A *reduct* of the whole set of conditional features $\mathbb{C}$is defined as the minimal subset of conditional attributes $R \subset \mathbb{C}$ such that, for a given decision attribute D, $\gamma_R(D) = \gamma_{\mathbb{C}}(D)$. This means that no attributescan be removed from the subset without affecting the dependency degree, formally:

$$\gamma_{R-\{a\}}(D) \neq \gamma_R(D), for\ all\ a \in R.$$

As for a given dataset many reducts may exist, the intersection of all reducts is called *core* and it is made of those features that cannot be eliminated without information loss.

3.3.2 Variable Precision Rough Set Theory

Although widely used for classification tasks and FS problems, RST presents some limits that can be summarized in the points below [19]:

- ✓ RST allows managing only objects which classification is completely correct or certain. This means that no degree of uncertain in classification is admitted, even if very frequently the information about data are only partial;
- ✓ RST assumes that the entire universe $\mathbb{U}$is composed only from data under consideration. Consequently, the conclusions derived from the model are applicable exclusively only to this set of elements.

Variable Precision Rough Set (VPRS) theory was introduced by Ziarko [19] as a generalization of the standard RST which surpasses the above depicted limits of RST, by admitting a certain degree of objects misclassification.

Let X and Y be non-empty subsets of a universe $\mathbb{U}$ of objects. VPRS introduces the measure of *relative degree of misclassification* $c(X,Y)$ of the set X with respect to the set Y calculated as:

$$c(X,Y) = 1 - \frac{|X \cap Y|}{|X|}$$

where $|\cdot|$ denotes the set cardinality. It can be observed that $c(X,Y) = 0$ if and only if $X \subseteq Y$.

Considering $0 \leq \beta < 0.5$ as the *admissible classification error*, it is possible to redefine all rough-set concepts. Also the inclusion relationship that can be redefined with a certain level of error β as:

$$X \subseteq_\beta Y \text{if and only if} c(X,Y) \leq \beta.$$

By replacing the standard inclusion relationship with this new generalized relation, the β-approximations and the β-regions can be formulated as:

$$\underline{P_\beta} X = \{x \in \mathbb{U} | [x]_P \subseteq_\beta X\} = \{x \in \mathbb{U} | c([x]_P, X) \leq \beta\}$$

$$\overline{P}_\beta X = \{x \in U | c([x]_P, X) \leq 1 - \beta\}$$

$$POS_{P,\beta}(D) = \bigcup_{X \in \mathbb{U}/IND(D)} \underline{P_\beta} X$$

$$NEG_{P,\beta}(D) = U - \bigcup_{X \in \mathbb{U}/IND(D)} \bar{P}_\beta X$$

$$BND_{P,\beta}(D) = \bigcup_{X \in \mathbb{U}/IND(D)} \bar{P}_\beta X - \bigcup_{X \in \mathbb{U}/IND(D)} \underline{P_\beta} X.$$

Let $P \subseteq \mathbb{C}$ be a subset of conditional features and $D \subseteq \mathbb{D}$ the decision feature, it is possible to measure the β-*dependency degree* as:

$$\gamma_{P,\beta}(D) = \frac{\left|POS_{P,\beta}(D)\right|}{|\mathbb{U}|}$$

where $|\cdot|$ denotes the cardinality of set. The standard RST definitions are obtained from the above equations setting $\beta = 0$.

Referring to the dataset in Table 3.I and considering $P = \{b, c\}$, $D = \{e\}$ and $\beta = 0.4$, the partitions of the universe created by P and D are:

$$\mathbb{U}/IND(P) = \big\{\{3\}, \{1,5\}, \{4\}, \{2,7,8\}, \{6\}\big\}$$

$$\mathbb{U}/IND(D) = \{\{1\},\{2,4,7\},\{3,5,6,8\}\}.$$

For each equivalence class $X \in \mathbb{U}/IND(P)$ and $Y \in \mathbb{U}/IND(D)$, the value of the relative degree of misclassification $c(X,Y)$ has to be calculated and it must be less than β in order to include the equivalence class Xin the β-positive region. Considering $X = \{2,7,8\}$, it can be obtained:

$$c(\{2,7,8\},\{1\}) = 1 - \frac{0}{3} = 1 > \beta$$

$$c(\{2,7,8\},\{2,4,7\}) = 1 - \frac{2}{3} = 0.33 < \beta$$

$$c(\{2,7,8\},\{3,5,6,8\}) = 1 - \frac{1}{3} = 0.66 > \beta$$

Only objects 2, 7 and 8 are included in the β-positive region because they belong to a β-subset of $\{2,4,7\}$. Repeating the same calculation for all subsets in $\mathbb{U}/IND(P)$, the β-positive region is obtained as:

$$POS_{P,\beta}(D) = \{2,3,4,6,7,8\}$$

and, consequently, the β-dependency degree is:

$$\gamma_{P,\beta}(D) = \frac{|\{2,3,4,6,7,8\}|}{|\{1,2,3,4,5,6,7,8\}|} = \frac{6}{8}$$

3.4 Feature Extraction by means of Time-Frequency Analysis

Before applying the FS methods, it is needed to construct a proper target dataset for knowledge discovery. This phase, usually called *Feature Extraction*, is carried out in different ways, according to the specific application, and aims to identify the important characteristics representing the context of interest. In the biomedical field, these features are often associated with a particular pathology, motion or state of the body.

In some cases, features are extracted directly from measurements or experts' judgments and organized in order to obtain the appropriate dataset. For more complex applications, datasets are built applying, to a set of recorded signals, different processing techniques.

The latter is the case of most biomedical applications, in which signals derived from biological processes are often analyzed both in the time and in the frequency domain. The methods applied in frequency domain analysis depend on the stationarity or non-stationarity of the signals. When signals are non-stationary, the traditional Fourier-based spectral analysis cannot be applied and time-frequency distributions should be preferred.

In this thesis, the signals recorded by the NIRS were taken into account for two applications (see Chapters 4 and 5). This technique allows measuring the changes in the concentration of the oxygenated (O_2Hb) and reduced (HHb) hemoglobin during active stimuli, such as breath-holding (BH) or hyperventilation (HYP). An example of NIRS signals is reported in Figure 3.7, showing the changes in the concentration of O_2Hb and HHb during BH, measured by the NIRS. As it can be seen in the figure, these signals rapidly change during the stimulus, requiring the use of time-frequency distributions for their management.

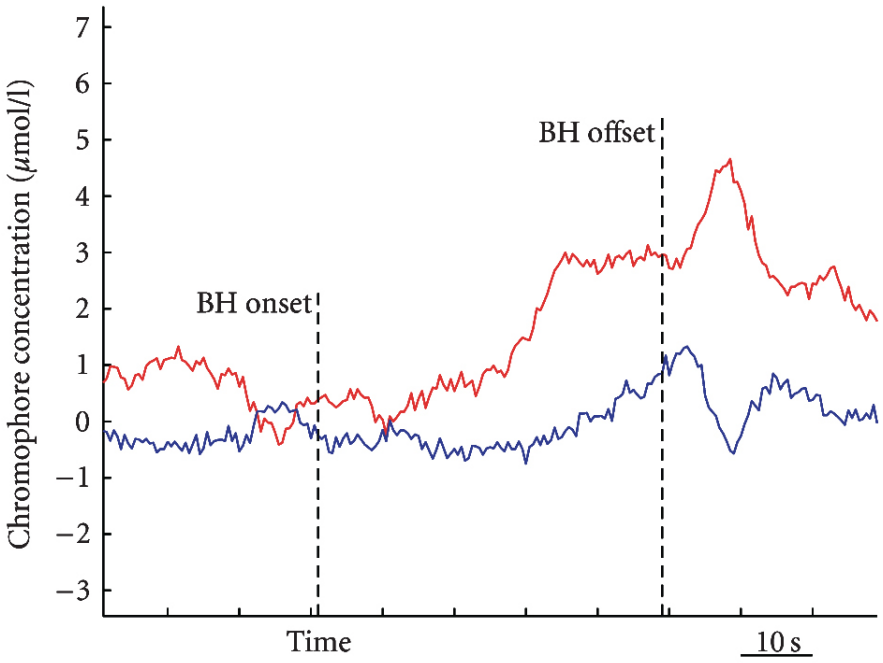

Figure 3.7 NIRS signals recorded on a healthy woman performing breath – holding. The red line represents the changes in the O_2Hb concentration signal, the blue line the HHb. The black vertical dashed lines mark the onset and offset of the breath - holding.

Specifically, a time-frequency distribution belonging to the *Cohen's class*$D_{xx}(t,f)$ was chosen in this thesis, defined as:

$$D_{xx}(t,f) = \iiint_{-\infty}^{+\infty} x\left(t' - \frac{\tau}{2}\right) x^*\left(t' + \frac{\tau}{2}\right) g(\tau,\theta) e^{-j2\pi\theta(t'-t)} e^{-j2\pi f\tau}\, dt'\, d\theta\, d\tau$$

where $x(t)$ is the signal under analysis, θ and τ are the frequency and time lags respectively, and $g(\tau,\theta)$ is the kernel function of the time-frequency distribution.

In particular, the *Choi-Williams transform* [20] was used, having a kernel expressed as:

$$g(\tau,\theta) = e^{-(\tau^2\theta^2/\sigma)}$$

where σ is a parameter influencing the kernel selectivity: a lower σ value will create a higher attenuation of interference terms. A Choi-Williams representation of the HHb signal during BH for a healthy subject is reported in Figure 3.8.

Moreover, the time-frequency *Squared Coherence Function* (SCF) between two signals x and y was analyzed, calculated basing on the Choi-Williams representation as:

$$SCF_{xy}(t,f) = \frac{\left|D_{xy}(t,f)\right|^2}{D_{xx}(t,f) \cdot D_{yy}(t,f)}$$

where $D_{xy}(t,f)$ is the cross time-frequency transform between the signals, and $D_{xx}(t,f)$ and $D_{yy}(t,f)$ are the time-frequency representations of the x and y signals, respectively. As the SCF is a quadratic function, it assumes only values between zero, if the two signals are totally uncorrelated, and one, if they are totally correlated. Figure 3.9 presents an example of time-frequency SCF between the two NIRS signals during BH related to a healthy subject.

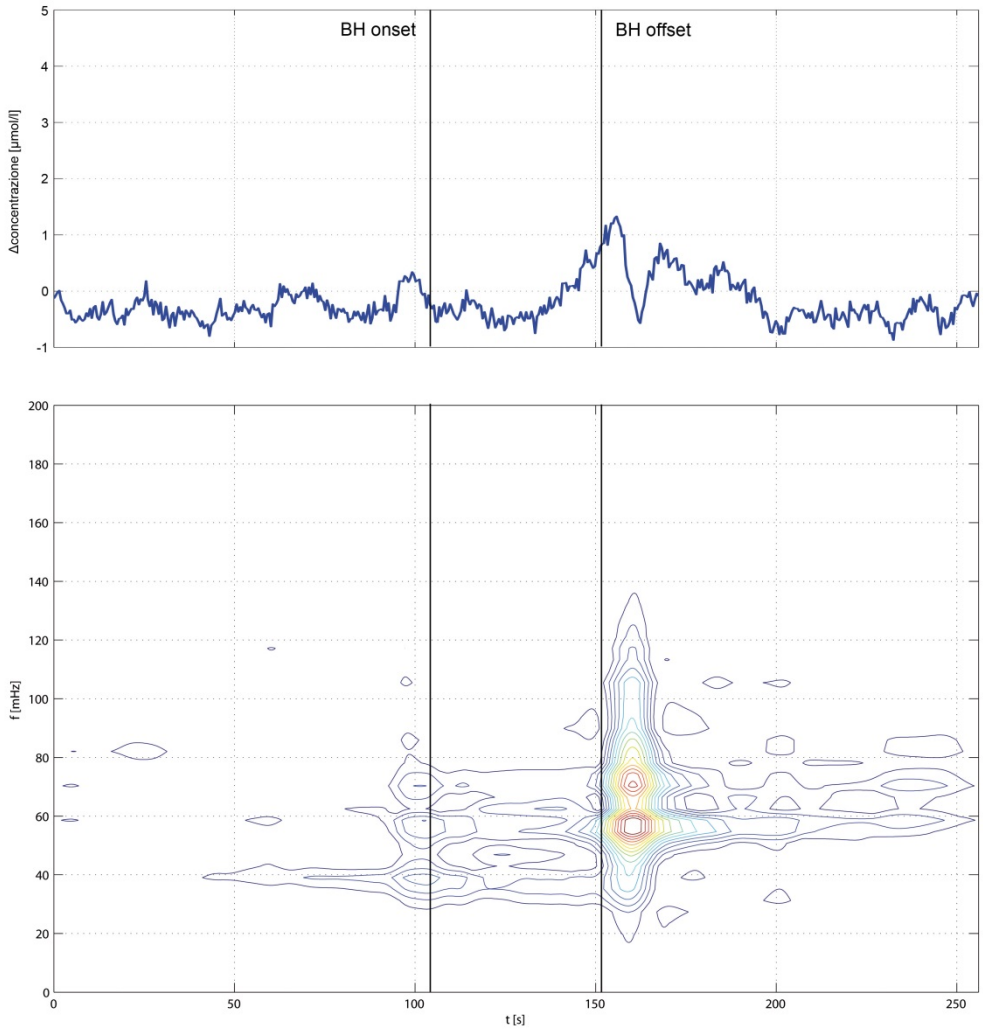

Figure 3.8 HHb concentration signal (upper panel) recorded on a healthy subject lasting 256 seconds with the BH in the middle of the analysis window. The onset and the offset of the event are marked by vertical lines. In the lower panel is showed the Choi-Williams distribution of the signal (σ=0.05) by 15-level curves. The graphs show that the NIRS signals become non-stationary as consequence of the active stimuli.

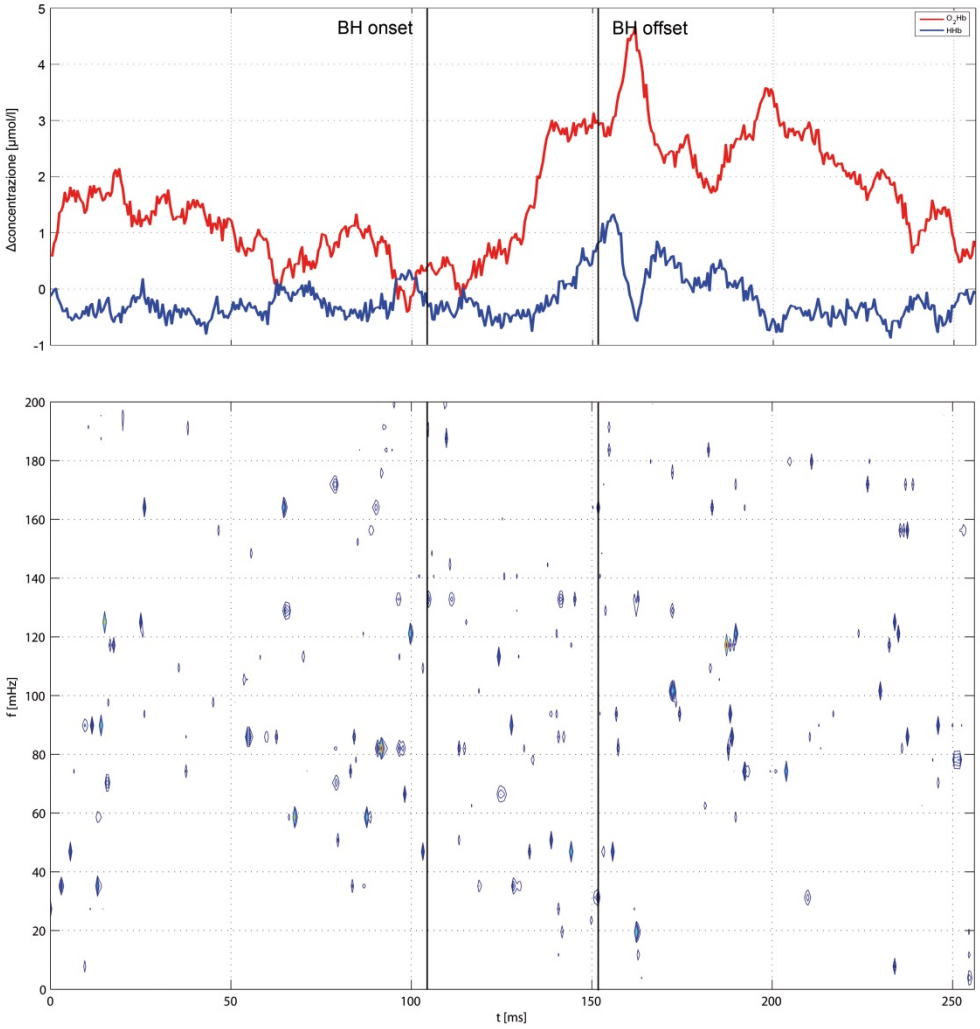

Figure 3.9 The HHb (blue line) and O_2Hb (red line) signals during BH (upper panel) related to a healthy subject. The onset and the offset of the event are marked by vertical lines. The lower panel shows the 15-level contour plot of the time-frequency SCF between the two signals.

3.5 Supervised FS Algorithms

In the following sections, five algorithms for automatic and non-linear FS used in this thesis are illustrated. This step of dimensionality reduction is of crucial importance in the KDD process in order to improve the quality of data, removing those variables that are irrelevant or sources of noise for the following phases of analysis.

3.5.1 QuickReduct Algorithm (QRA)

QuickReduct Algorithm (QRA), introduced by Shen and Chouchoulas [21], is a basic tool for solving reduct search problems without generating all the possible subsets of features. It is based on a forward heuristic selection using the Rough Set concept of *dependency degree* for measuring the consistency of a dataset.

The algorithm starts from an empty subset of features and adds to it the best attributes, one at a time, until a stopping criterion is satisfied. Specifically, in each iteration, the attribute producing the largest increase in the dependency degree is added to the current subset. As the goal of QRA is to find a subset of features with the same dependency degree of the whole set of attributes, the search stops when this condition is achieved. The pseudo-code of QRA is depicted in Figure 3.10.

```
QUICKREDUCT(C,D)
Input: C -> set of all conditional
features
       D -> set of decision features
Output:  R -> feature subset
(1) R <- {}
(2) while γR(D) ≠ γC(D)
(3)     T <- R
(4)     foreach x ∈ (C - R)
(5)         if γR∪{x}(D) > γT(D)
(6)             T <- R ∪ {x}
(7)     R <- T
(8) return R
```

Figure 3.10 *QuickReduct Algorithm* (QRA) pseudo-code [22].

3.5.2 Entropy-Based Reduction (EBR)

Entropy-Based Reduction (EBR), formulated by Jensen and Shen [23], is based on the measure of the *conditional information entropy* $H(D|A)$ produced by an attribute A with respect to the decision feature D. It evaluates the information content generated by an information source [24] and it is calculated as:

$$H(D|A) = -\sum_{j=1}^{m}\left(p(a_j)\sum_{i=1}^{n} p(c_i|a_j)\, log_2\, p(c_i|a_j)\right) \quad (3.2)$$

where $a_1 \dots a_m$ and $c_1 \dots c_n$ are the values of attributes A and D respectively, $p(a_j)$ is the probability that the value a_j occurs and $p(c_i|a_j)$ is the conditional probability of a_j given c_i. Equation (3.2) can be easily extended from one conditional attribute to the whole set of attributes, returning the whole information content.

The structure of EBR algorithm is similar to the one of QRA: in each iteration, it adds to the current subset those features producing the highest decrease of entropy. The reduct search stops when the variable subset reaches the same entropy of all available attributes (zero for consistent dataset). In Figure 3.11 it is showed the EBR pseudo-code.

```
EBR(C,D)
Input: C -> set of all conditional features
       D -> set of decision features
Output: R -> feature subset
(1) R <- {}
(2) while H(D|R) ≠ H(D|C)
(3)     T <- R
(4)     foreach x ∈ (C - R)
(5)         if H(D|R ∪ {x}) < H(D|T)
(6)             T <- R ∪ {x}
(7)     R <- T
(8) return R
```

Figure 3.11 *Entropy-Based Reduct* (EBR) algorithm pseudo-code [22].

3.5.3 GenRSAR

Genetic Algorithms (GAs), introduced by Holland [25] and belonging to the EC, are a class of metaheuristics tending to mime the natural evolutionary process of species. Because of their intrinsic characteristics, GAs easily adapt to FS problems. In fact, as the chromosome is represented as a binary string of bits, in FS each bit is associated to a specific feature and the chromosome length is equal to the number of available features. A one or zero in a certain position of

the string indicates respectively if the feature is included or not in the current subset.

In [16], Jensen and Shen introduce the *GenRSAR* (*Genetic Rough Set Attribute Reduction*), a genetic-based algorithm that uses the evolutionary search strategy in order to find rough set reducts. The algorithm follows the standard GA procedure in which the fitness function express the suitability of the feature subset as:

$$fitness(R) = \gamma_R(D) * \frac{|\mathbb{C}| - |R|}{|\mathbb{C}|}$$

Where $\gamma_R(D)$ is the dependency degree of the subset $R$with respect to the classification D, and $\mathbb{C}$ is the whole set of conditional features. The pseudo-code of GenRSAR is showed in Figure 3.12.

```
GenRSAR (C,D)
Input: C -> set of all conditional features
       D -> set of decision features
Output: R -> feat. subset
(1) Creare an initial population
(2) while stopping condition not true
(3)    Select parents
(4)    Perform crossover
(5)    Mutate offspring
(6)    Select new population
(7) return R (best solution)
```

Figure 3.12 *Genetic Rough Set Attribute Reduction* (GenRSAR) algorithm pseudo-code.

3.5.4 AntRSAR

Ant Colony Optimization (ACO) is a methodology belonging to the SI algorithms inspired to the social behavior of ants for finding the best route in the direction of food.

This methodology, as GAs, is easily adaptable for FS problems. As ACO requires a problem to be represented as a graph, in FS nodes are associated to features (one node for each attribute) and edges between nodes may denote the choice of the next feature.

In [16] a RST-based ACO, called *Ant Rough Set Attribute Reduction* (*AntRSAR*), is proposed. In the algorithm, the dependency degree is used to verify

the construction of a completed solution for each ant. This means that an ant stops building a feature subset when it reaches the same *dependency degree* of the whole attribute set. As for the desirability measure, that is the probability to move from a node to another one, the *conditional information entropy* is used. The number of ants is set equal to the number of features, with each ant starting on a different feature. Figure 3.13 shows the pseudo-code of AntRSAR algorithm.

```
AntRSAR (C,D)
Input: C -> set of all conditional features
       D -> set of decision features
Output: R -> feat. subset
(1) Place an ant on each inizial node
(2) while stopping condition not true
(3)    foreach ant
(4)          Construct a path (solution)
(5)    Evaporate pheromone
(6)    Update pheromone
(7) return R (best solution)
```

Figure 3.13 *Ant Rough Set Attribute Reduction* (AntRSAR) algorithm pseudo-code.

3.5.5 Improved QuickReduct Algorithm (IQRA)

Deriving from RST, also QRA presents some limitations, mainly due to the assumption about the monotonicity of dependency degree, which has to increase in each iteration. If this condition is not satisfied, a random choice of the next feature is performed during the algorithm, leading to obtain a reduct with more attributes. Moreover, QRA ignores the redundancy in the dataset: the objects included in the positive region in an intermediate iteration will not add any more knowledge in the remaining iterations [26], and could be removed from the current dataset.

All this reasons leads researchers to modify the standard algorithm using different strategies, in order to improve the FS and consequently the DM phase. One of these approaches, based on VPRS theory, was proposed by Prasad and Rao [26]. In this study, the authors present the Improved QuickReduct Algorithm (IQRA), which takes into account redundant elements by deleting them from the analyzed dataset.

```
IMPROVED QUICKREDUCT(C,D)
Input: C -> set of all conditional features
       D -> set of decision features
       U -> object dataset
Output: R -> feature subset
(1) R <- {}
(2) Count=0
(3) DT=U
(4) while γ_{R,1}(D) ≠ γ_{C,1}(D)
(5)     AvailableSet=C-R
(6)     β=1
(7)     ε=0.1
(8)     T <- R
(9)     foreach x ∈ AvailableSet
(10)        if γ_{R∪{x},β}(D) > γ_{T,β}(D)
(11)            T <- R ∪ {x}
(12)    If T=R
(13)        β = β − ε
(14)        If β ⩾ 0.5
(15)            goto (8)
(16)        else
(17)            R=R ∪ {first attribute in AvailableSet}
(18)    else
(19)        R <- T
(20)    POSPARTIAL = POS_{R,1}(D)
(21)    DT = DT - POSPARTIAL
(22)    Count = Count + |POSPARTIAL|
(23)    γ_{R,1}(D) = Count/|U|
(24)return R
```

Figure 3.14 *Improved QuickReduct Algorithm* (IQRA) pseudo-code.

IQRA basic idea is similar to QRA: it starts with an empty features subset and adds, at every iteration, those attributes inducing the greatest increment in the dependency degree. If there is no increase in the dependency degree during an iteration, the VPRS concept of admissible classification error β is taken into account. In such a situation the β-dependency degree is calculated starting from β equal to 1 and decreasing its value until it cannot be reduced any more. In the latter case, the first available feature is included in the current subset. After the inclusion of a new attribute in the subset, the positive region with β equal to 1 is calculated and the objects belonging to this set are removed from the dataset. These elements, in fact, will result redundant for the next iterations. The algorithm stops when the dependency degree of the feature subset considering β

equal to 1 reaches the maximum value. The IQRA pseudo-code is depicted in Figure 3.14.

3.6 Intelligent Systems for Classification

Once the phase of data cleaning is performed during the KDD process, different intelligent tools are applied in order to discover new knowledge from data. In this study, two CI techniques are used for classification with two different aims, basing on the context and the specific application. In some cases the intelligent systems are employed directly for DM and extraction of knowledge from data (see Chapters 4 and 5). The basic idea in this kind of applications is that a good procedure of FS allows removing redundant and irrelevant features so that the reduced subset provides the same quality of learning than the original set [27] or even improve it. In other applications, these intelligent approaches can be used as intermediate step for further phases of processing and interpretation of signals and images (as in Chapter 6).

3.6.1 Feed-Forward Neural Networks (FFNNs)

A Feed-Forward Neural Networks (FFNN) is composed by several ANs connected among them and organized in layers, in order to be able to solve more complex problems than the single perceptron.

A FFNN is made of an input layer receiving information from outside, a certain number of hidden layers elaborating the information, and an output layer that returns the result of the learning process to the user. The neurons in each layer are connected only with the ones in the following layer, allowing the information to move only from input to output, and not backwards. A typical 2-layers FFNN architecture is showed in Figure 3.15 (usually the input layer is not included in the count of network layers because it does not do any processing [28]).

The number of hidden layers needed to solve a specific problem is related to the complexity of the analyzed system. In particular, one hidden layer allows solving linear problems, while more complex solutions are achievable using a higher number of layers. A graphical interpretation of the role of hidden layers in solving complex problems is showed in Figure 3.16 for a two-dimensional input space.

Different learning algorithms are available for training a FFNN. In this thesis, the *back-propagation algorithm* is used for all applications. This methodol-

ogy employs the sum of squared errors for measuring the learning performance of the network. This means that the error between the current output and the desired output is calculated and propagated backward to adjust the weights connecting the layers.

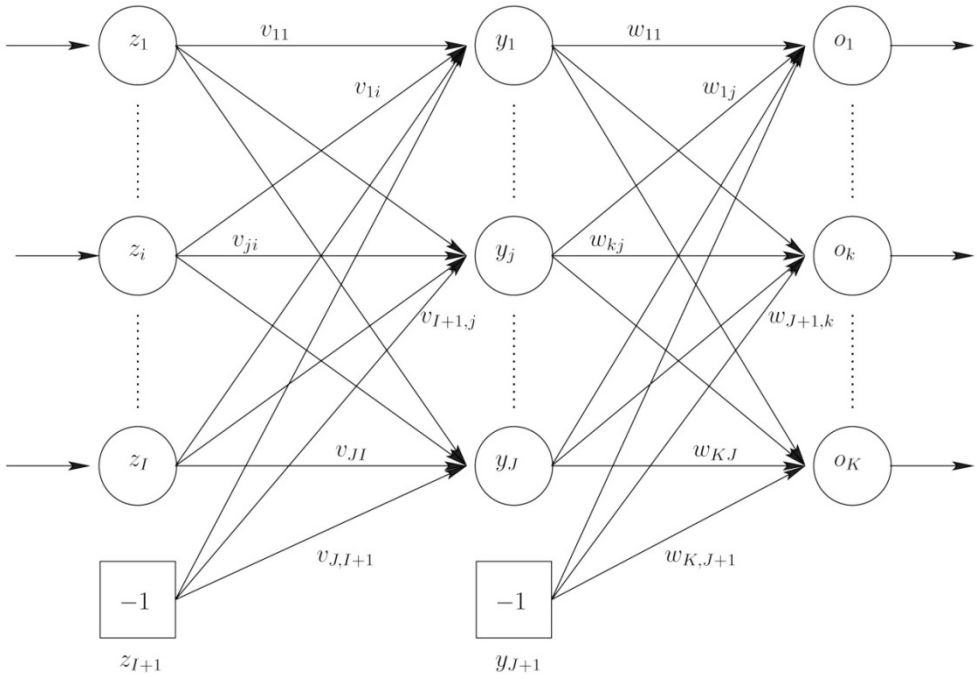

Figure 3.15 A typical 2-layers *Feed-Forward Neural Network* (FFNN) architecture made of one input layer, one hidden layer and one output layer [2].

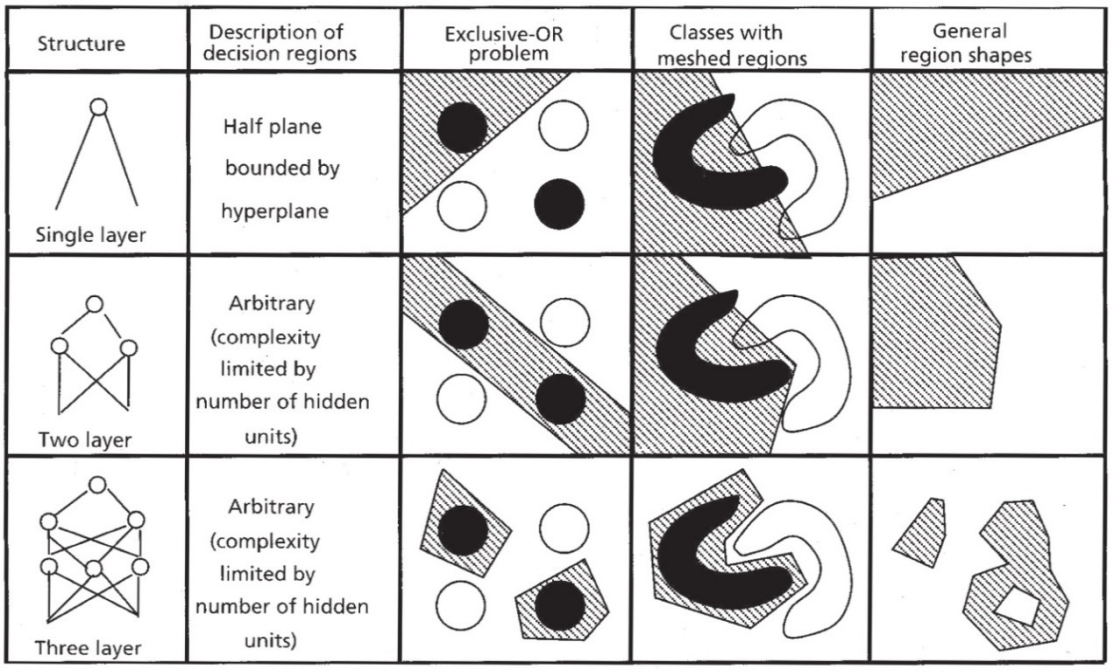

Figure 3.16 Graphical interpretation of the role of hidden layers for a two-dimensional input space [29].

3.6.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a method for supervised learning widely used for binary classification and regression.

Given a set of training examples, SVM projects the samples in a new space usually with higher dimensionality, in which the objects belonging to the two classes are linearly separable. In fact, in the new space, SVM finds a separating hyperplane able to maximize its distance from the nearest points of the two classes. These points are called *support vectors* and their distance from the hyperplane is defined as *margin*. In Figure 3.17 it is depicted a 2-D example of SVM classifier and it can be observed that the support vectors are the most critical elements for the finding of the separation hyperplane, as the margin distance is related only to these points.

For complex problems, linear separation could be impossible to be obtained directly. In these situations, a non-linear function called *kernel* is used to map data from the initial space to another space (usually with higher dimensionality) in which the linear separation between the two classes can be reached. Several kernel functions are available for solving complex classification tasks, such as linear, polynomial, or radial basis function, according to the application.

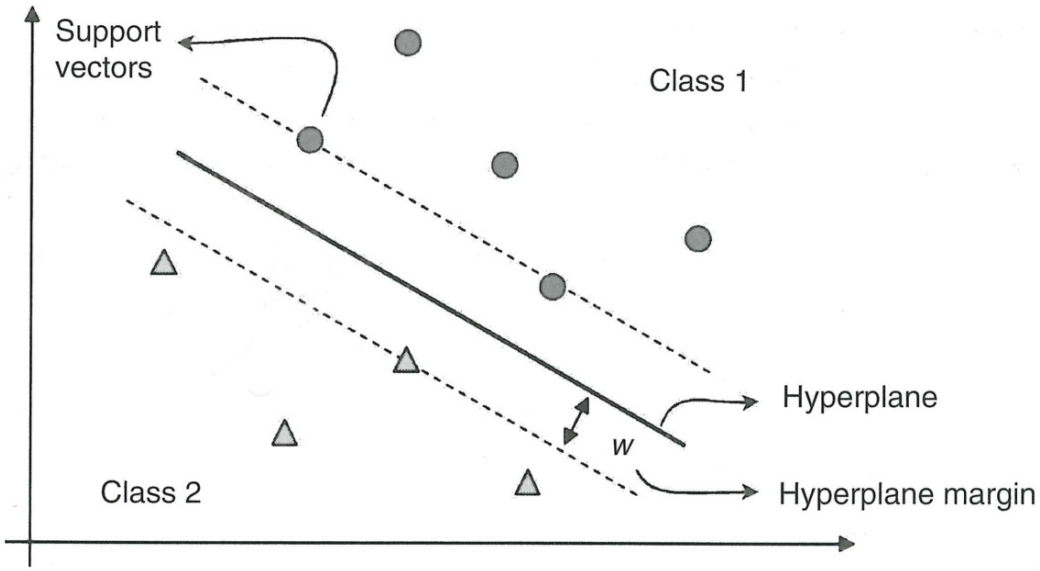

Figure 3.17 2-D example of a *Support Vector Machine* (SVM) classifier.

References

1. Venayagamoorthy, G.K.K., *A successful interdisciplinary course on coputational intelligence.* Computational Intelligence Magazine, IEEE, 2009. **4**(1): p. 14-23.
2. Engelbrecht, A.P., *Computational Intelligence: An Introduction*. 2007: Wiley Publishing. 628.
3. Kerre, E.E. and M. Nachtegael, *Fuzzy Techniques in Image Processing*. 2000: Physica-Verlag HD.
4. Jiang, J., P. Trundle, and J. Ren, *Medical image analysis with artificial neural networks.* Computerized Medical Imaging and Graphics, 2010. **34**(8): p. 617-631.
5. Rao, V.S.H. and M.N. Kumar, *A new intelligence-based approach for computer-aided diagnosis of dengue fever.* IEEE Transactions on Information Technology in Biomedicine, 2012. **16**(1): p. 112-118.
6. Negnevitsky, M., *Artificial Intelligence: A Guide to Intelligent Systems*. 2005: Addison-Wesley.
7. Mamdani, E.H. and S. Assilian, *An experiment in linguistic synthesis with a fuzzy logic controller.* International Journal of Man-Machine Studies, 1975. **7**(1): p. 1-13.
8. Pawlak, Z., *Rough sets.* International Journal of Parallel Programming, 1982. **11**(5): p. 341-356.
9. Drwal, G., *Rough, and Fuzzy-Rough Classification Methods Implemented in RClass System*, in *Rough Sets and Current Trends in Computing*, W. Ziarko and Y. Yao, Editors. 2001, Springer Berlin Heidelberg. p. 152-159.
10. Shen, Q. and R. Jensen, *Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring.* Pattern Recognition, 2004. **37**(7): p. 1351-1363.
11. Li, F., M. Ye, and X. Chen, *An extension to Rough c-means clustering based on decision-theoretic Rough Sets model.* International Journal of Approximate Reasoning, 2013.
12. Chu, X.Z., et al., *An expert system using rough sets theory and self-organizing maps to design space exploration of complex products.* Expert Systems with Applications, 2010. **37**(11): p. 7364-7372.
13. Fan, T.F., D.R. Liu, and G.H. Tzeng, *Rough set-based logics for multicriteria decision analysis.* European Journal of Operational Research, 2007. **182**(1): p. 340-355.
14. AboulElla, H., et al., *Rough Sets in Medical Imaging: Foundations and Trends*, in *Computational Intelligence in Medical Imaging*, G. Schaefer, A. Hassanien, and J. Jiang, Editors. 2009, Chapman and Hall/CRC. p. 47-87.
15. Banerjee, M., S. Mitra, and A. Anand, *Feature Selection Using Rough Sets*, in *Multi-Objective Machine Learning*, Y. Jin, Editor 2006, Springer: Heidelberg. p. 3-20.

16. Jensen, R. and Q. Shen. *Finding Rough Set Reducts with Ant Colony Optimization*. in *Proceedings of the 2003 UK Workshop on Computational Intelligence*. 2003. Bristol.
17. Jensen, R. and Q. Shen, *Fuzzy-Rough Sets Assisted Attribute Selection.* Fuzzy Systems, IEEE Transactions on, 2007. **15**(1): p. 73-89.
18. Schaefer, G., A. Hassanien, and J. Jiang, *Computational Intelligence in Medical Imaging: Techniques and Applications*. 2009: Taylor & Francis.
19. Ziarko, W., *Variable precision rough set model.* J. Comput. Syst. Sci., 1993. **46**(1): p. 39-59.
20. Cohen, L., *Time-frequency distributions - A review.* Proceedings of the IEEE, 1989. **77**(7): p. 941-981.
21. Shen, Q. and A. Chouchoulas, *A modular approach to generating fuzzy rules with reduced attributes for the monitoring of complex systems.* Engineering Applications of Artificial Intelligence, 2000. **13**(3): p. 263-278.
22. Jensen, R. and Q. Shen, *Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches*. 2008, Hoboken, NJ: Wiley-IEEE Press. 340.
23. Jensen, R. and Q. Shen, *A Rough Set-Aided System for Sorting WWW Bookmarks*, in *Proceedings of the First Asia-Pacific Conference on Web Intelligence: Research and Development*2001, Springer-Verlag. p. 95-105.
24. Małyszko, D. and J. Stepaniuk, *Adaptive multilevel rough entropy evolutionary thresholding.* Information Sciences, 2010. **180**(7): p. 1138-1158.
25. Holland, J.H., *Adaptation in natural and artificial systems*. 1992: MIT Press. 211.
26. Prasad, P. and C. Rao, *IQuickReduct: An Improvement to Quick Reduct Algorithm*, in *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, H. Sakai, et al., Editors. 2009, Springer Berlin / Heidelberg. p. 152-159.
27. Chen, Y.M., et al., *A rough set approach to feature selection based on power set tree.* Knowledge-Based Systems, 2011. **24**(2): p. 275-281.
28. Begg, R., D.T.H. Lai, and M. Palaniswami, *Computational Intelligence in Biomedical Engineering*. 2007: Taylor & Francis.
29. Jain, A.K., M. Jianchang, and K.M. Mohiuddin, *Artificial neural networks: a tutorial.* Computer, 1996. **29**(3): p. 31-44.

Chapter 4: Statistical vs Non-Statistical FS Methods

4.1 Introduction

The first aim of this study was to compare statistical approaches for FS with advanced ones, based on RST. In particular, we want to assess if the classical methods based on statistics are able to manage small and unbalanced biomedical datasets, characterized by non-linearity between features and classification. For this purpose, we used a dataset related to the cerebral hemodynamics of migraine sufferers, based on Near-InfraRed Spectroscopy (NIRS) signals.

Migraine is a neurological disorder correlating with an increased risk of subclinical cerebral vascular lesions [1]. Epidemiological studies showed that migraineurs are prone to an increased risk of vascular accidents [2], and this lead many researchers to consider this pathology as a systemic vasculopathy [3]. The association between migraine and impaired cerebral autoregulation or vasomotor tone has widely been investigated and assessed [4-6].

However, there is a clear difference in the cardiovascular and cerebrovascular risk associated to the different types of migraine. Subjects suffering from migraine with aura (MwA) showed greater impairments with respect to subjects suffering from migraine without aura (MwoA) [1, 2, 4, 6]. Aura is a specific disturbance associated with migraine that can cause visual, speech, or perceptional impairments. Given the correlation between migraine and vascular disorders, migraine sufferers usually undergo assessments of the cerebrovascular

status. The accurate assessment of the cerebrovascular reactivity can be of paramount importance for the onset of a personalized and proper therapy.

Near-infrared spectroscopy (NIRS) is a non-invasive, real-time, and cost-effective monitoring technique for the assessment of the cerebral autoregulation of subjects [4]. Infrared light is injected into the skull and the changes in the concentration of oxygenated (O_2Hb) and deoxygenated (or reduced) (HHb) hemoglobin are instantaneously measured. A deeper description of the NIRS System is provided in [7].

The assessment of cerebral vasomotor reactivity (i.e., the arteries' capacity of compensating systemic blood pressure alterations) is of primary importance to evaluate the overall status of the artery bed. Usually, the active maneuvers like breath-holding (BH), hyperventilation (HYP), or Valsalva, are performed during the monitoring phase in order to assess cerebral autoregulation and vasomotor reactivity [4, 8, 9]. Such maneuvers are easy to perform and are safe also for pathological subjects. Specifically, BH determines the increase of the carbon dioxide in the blood, causing a vasodilatation of the cerebral vessels, whereas HYP triggers vasoconstriction because of the increase of oxygen in the blood. Therefore, overall, the NIRS is a suitable system for long-term, bedside, or home monitoring and assessment. The use of NIRS for the assessment of migraineurs is gaining clinical importance. Recently, Watanabe et al. used NIRS to monitor the hemodynamical changes occurring during a migraine attack after the administration of sumatripan [10]. Also, Viola et al. studied the pathophysiology of prolonged migraine attacks by monitoring the cerebral oxygenation [11].

The accurate assessment of the cerebral autoregulation of migraineurs is complicated by the fact that NIRS concentration signals are usually considered as nonstationary when recorded during vaso-active maneuvers (i.e., breath-holding and hyperventilation). In 2000, Obrig et al. [12] studied the spontaneous low frequency oscillations of cerebral hemodynamics and metabolism in adult human head by using NIRS. Though conducted on healthy volunteers, this study introduced the possibility of frequency-derived parameters used to assess cerebral autoregulation. It is known that cerebral hemodynamic signals have a power spectrum essentially consisting of two different bands [13]:

- ✓ A *very low frequency* band (VLF – also called B-waves) that reflects the long-term autoregulation. At brain level, VLFs are thought to be generated by brain stem nuclei, which modulate the lumen of the small intracerebral vessels. In humans, the VLF is usually comprised between 20 and 40 mHz.

- ✓ A *low frequency* band (LF – also called M-waves) that is common to most mammalians. Such waves reflect the systemic oscillations of the arterial blood pressure and are modulated by the sympathetic system activity. LFs spans from about 40 to 140 mHz.

An example of NIRS signals recorded during the BH (panel A) and HYP (panel B) of a healthy subject is shown in Figure 4.1.

In this study, we present the comparison of the performances of the two methods for FS applied to a dataset of features describing the time and frequency changes of the hemoglobin (both in its oxygenated and reduced form) concentrations as measured by NIRS in a population of women suffering from MwA and MwoA.

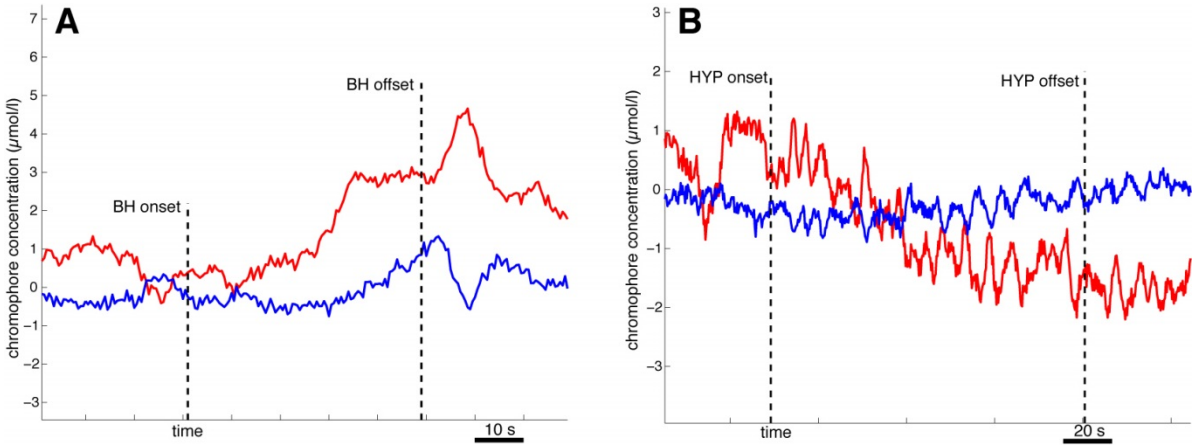

Figure 4.1 NIRS signals recorded on a healthy woman performing breath – holding (A) and hyperventilation (B). The red line represents the changes in the O_2Hb concentration signal, the blue line the HHb. The black vertical dashed lines mark the onset and offset of the breath - holding (A) and hyperventilation (B).

4.2 Methodology

4.2.1 Demographics

The study involved a total of 80 subjects, divided in 3 groups based on pathology: 15 healthy subjects were controls (Contr) (age: 29.2 ± 8.5), 14 women suffered from MwoA (age: 44.4 ± 9.7) and 51 women from MwA (age: 38.0 ± 12.1). Migraine with and without aura was diagnosed according to the criteria of the International Headache Society [14]. Migraine subjects were tested in their interictal periods (*i.e.* when they were free of pain).

The study received the approval from the Institutional Review Board of Gradenigo Hospital of Turin (Italy), where all the NIRS signals were acquired. All the subjects were instructed about the purposes of the study and signed an informed consent prior to undergoing tests.

4.2.2 Data Acquisition & Experimental Protocol

The recordings were performed in a quiet room, with dimmed lighting and a constant temperature of 24°-25°C. The subjects were lying in a supine position, with their eyes closed and breathing room air. Before and after active maneuvers, they performed a resting period of 120 seconds.

A commercial NIRS device (NIRO 300, Hammamatsu Photonics, Australia) was used to perform the experiments. The light source was placed on the forehead left side, about 2 cm alongside the midline and 3 cm above the supraorbital ridge. The distance between the source and the receiver was equal to 5 cm, and the differential path length factor was set to 5.97 [15]. Both the changes in the concentration of O_2Hb and HHb were acquired using a sampling frequency equal to 2 Hz.

4.2.3 Feature Extraction

As indicated in Figure 4.1, the vasodilatation due to BH produces an increase in the O_2Hb and a decrease in the HHb concentration (Figure 4.1(A)), whereas the vasoconstriction induced by HYP causes a decrease in the O_2Hb and an increase in the HHb concentration (Figure 4.1(B)). Moreover, in Figure 4.1(B) a periodic trend is visible on both the signals reflecting the respiratory rate.

Because of the rapid changes in NIRS signals, especially during active stimuli, the signals must be considered non-stationary and the Choi-William distribution was preferred for their analysis (§3.4). The value of σ was set equal to 0.5 for both signals, because this value already proved effective in the analysis of several biological signals. A Choi-Williams representation of the HHb NIRS signal during BH for a healthy subject is reported in Figure 4.2.

Moreover, the time-frequency *Squared Coherence Function* (SCF), between the concentration signals of O_2Hb and HHb, was computed basing on the Choi-Williams representation of the two signals. Figure 4.3 represents an example of time-frequency SCF between the two NIRS signals during BH relative to a healthy subject.

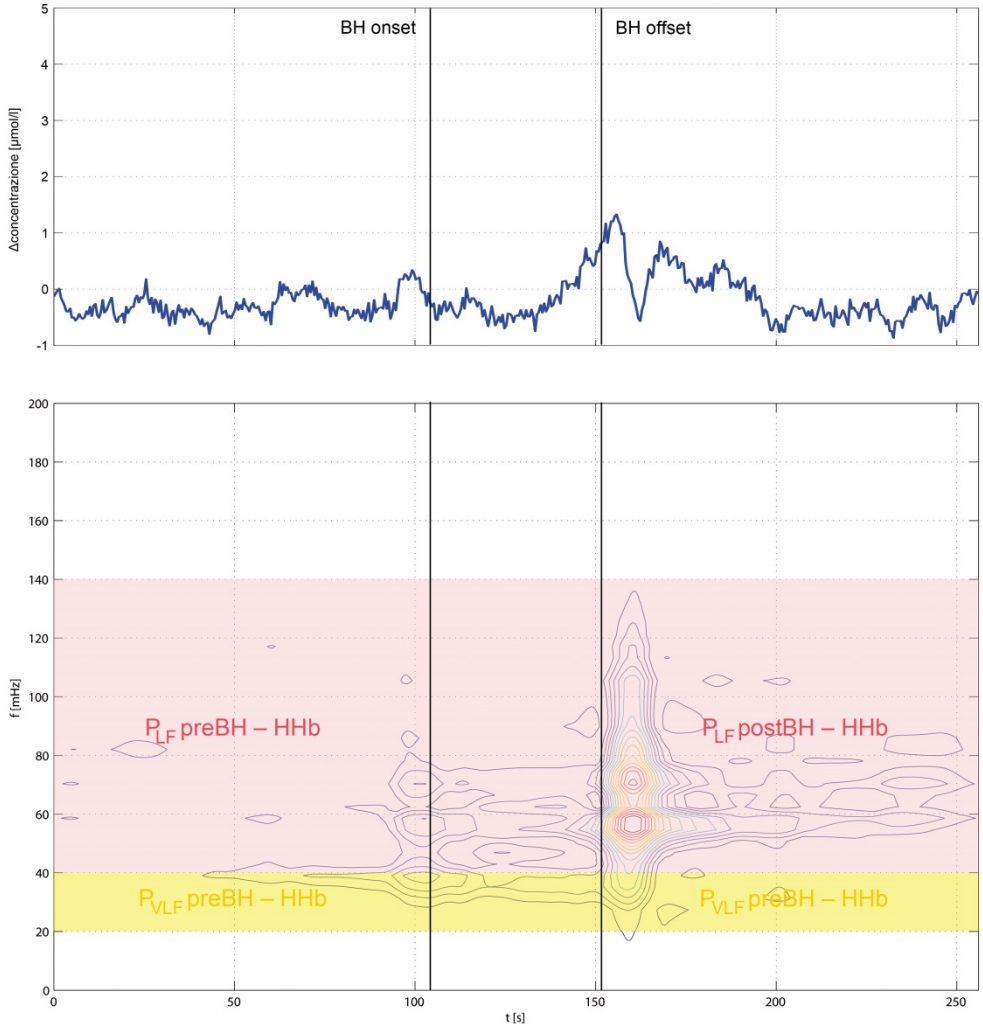

Figure 4.2 HHb concentration signal (upper panel) recorded on a healthy subject lasting 256 seconds with the BH in the middle of the analysis window. The onset and the offset of the event are marked by vertical lines. In the lower panel is showed the Choi-Williams distribution of the signal (σ=0.05) by 15-level curves. The yellow zone represents the VLF band (20-40 mHz) while the pink one indicates the LF band (40-140 mHz).The graphs show that the NIRS signals become nonstationary as consequence of the active stimuli.

We considered signals lasting 256 seconds, with the active maneuvers (BH and HYP) in the middle of the analysis window (see Figure 4.1). Hence, spectral resolution was better than 4 mHz, which has been shown to be a good value to clearly separate the two frequency bands of interest [4, 12].

Before the calculation of the time-frequency distributions, all the signals were converted to their analytical representation with zero mean and no trend. Trends were removed by using high-pass filtering (Chebychev filter, with ripple in the stopband and cutoff frequency equal to 15 mHz).

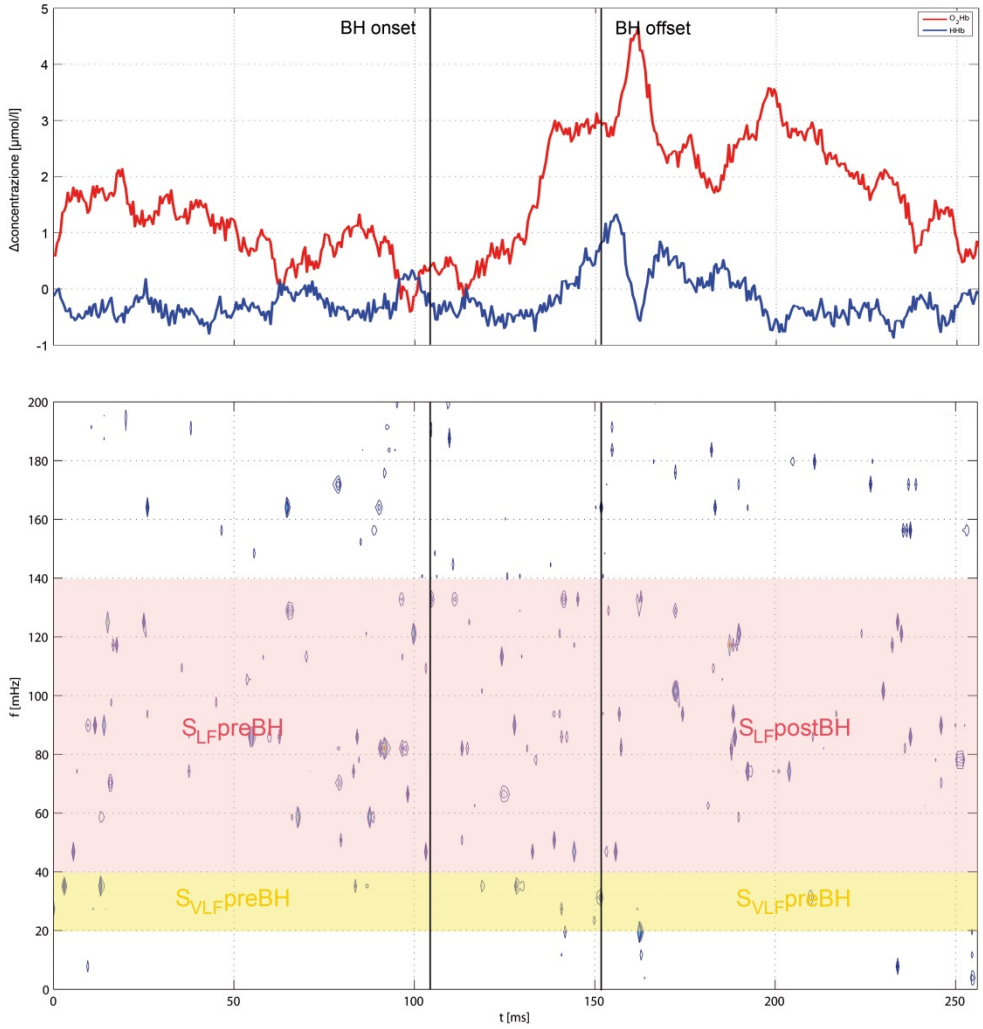

Figure 4.3 The HHb (blue line) and O_2Hb (red line) signals during BH (upper panel) relative to a healthy subject. The onset and the offset of the event are marked by vertical lines. The lower panel shows the 15-level contour plot of the time-frequency SCF between the two signals. The yellow zone represents the VLF band (20-40 mHz) while the pink one indicates the LF band (40-140 mHz).

The time–frequency distributions, concerning both Choi-Williams transforms of the two signals and the SCF, were analyzed in two specific bands, VLF and LF, before and after BH and HYP. The percentage of signal power in the two bands (referred to the total power of the signal) was calculated before and after each event. In this way, the following 24 variables deriving from the time-frequency representations were measured:

- ✓ the HHb and O_2Hb power in the VLF and LF bands (P_{VLF} and P_{LF}), before and after BH (for a total of 8 variables);
- ✓ the HHb and O_2Hb power in the VLF and LF bands (P_{VLF} and P_{LF}), before and after HYP (for a total of 8 variable);

- ✓ the O_2Hb and HHb SCF value in the two bands (S_{VLF} and S_{LF}), before and after BH and HYP (for a total of 8 variables).

Moreover, two additional variables, derived from the analysis in the time domain, were extracted: the Breath-Holding Index (BHI) for HHb and O_2Hb signals. The BHI for O_2Hb was calculated as:

$$BHI_{O2Hb} = \frac{[O2Hb]_{BH} - [O2Hb]_{BASE}}{D_{BH}}$$

where the numerator represents the variation of the O_2Hb concentration as effect of BH, while denominator is the BH duration. Similarly, it is possible to calculate the BHI_{HHb}. A deeper description about the methods for signals' analysis used in this study can be found in [7].

The 26 final variables, calculated as described above, are reported by the first column of Table 4.I.

4.2.4 Feature Selection

Two methods were chosen to perform FS. The first approach, based on the RST, is the QRA (§3.5.1) while the second methodology, based on a statistical linear model of data, is the ANalysis Of VAriance (ANOVA).

ANOVA allows comparing two or more set of data, by analyzing the variance within and between data groups. In this application, the one-way ANOVA analysis was performed using the pathology as an independent variable and the 26 parameters for each subject as dependent variables, one at a time. Then, the variables with a p-value greater than 5% were neglected for the study.

Both procedures were implemented in MATLAB environment.

4.2.5 Classification

In this study, two different FS strategies were implemented and their performances were compared using FFNNs (§3.6.1).

Specifically, three different networks were built, using as inputs respectively all the 26 attributes, the variables selected by QRA and those resulted from the ANOVA analysis.

About the FFNN structure, only one hidden layer was chosen, with a number of neurons approximately equal to half the input neurons. As for the neuron activation functions, the logarithmic sigmoid function was used for the hidden layer and the linear function for the output layer. Back-propagation was chosen as the learning algorithm and the mean squared error was used as per-

formance function. The initial values of interconnection weights were set randomly. A schematic representation of the three FFNNs is reported in Figure 4.4.

The FFNNs were implemented by means of the Neural Network toolbox provided in Matlab environment that uses 60% of input data randomly selected as training set while the complete dataset was used as test set. As the sample numerosity for each class was non-homogeneous, this could influence the training step. As a matter of fact, it is possible that some classes were less represented than others in the sample used for the neural network training. This aspect, beyond our control, obviously influences the FFNN final performances. To avoid this problem, each FFNN was repeated ten times and the best performance is reported as result.

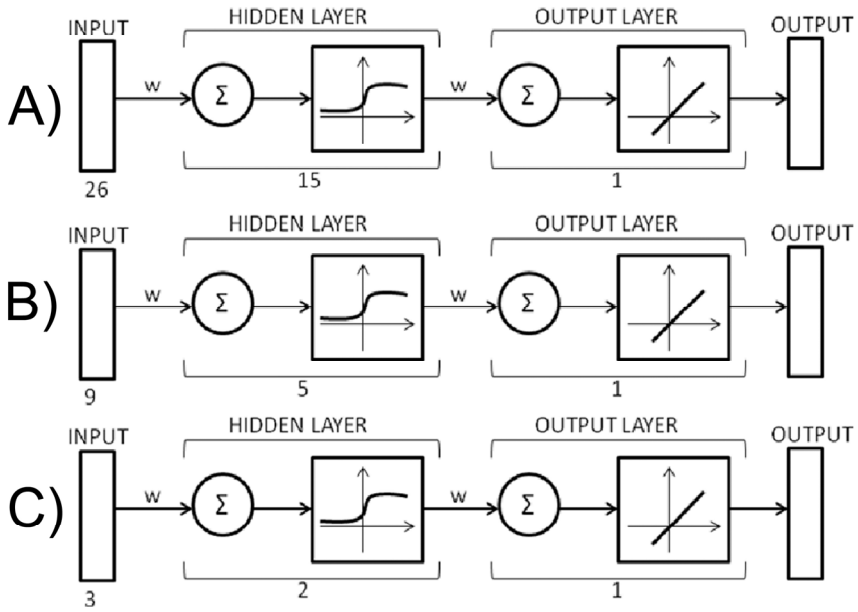

Figure 4.4 Schematic representation of FFNNs employed in order to test the two feature selection strategies. A) FFNN using the whole set of 26 available features as input data. B) FFNN using the QRA subset of features (9 parameters) as input data. C) FFNN using the ANOVA subset of features (3 parameters) as input data. All three networks consist of a hidden layer composed of a number of neurons that are variable according to the number of input features and with logarithmic sigmoid activation function. The output layer is made up of only one neuron with a linear activation function.

4.3 Results

In this application, two FS strategies were applied to a dataset related to migraine pathology. The dataset was made of 80 subjects characterized by 26 features listed in the first column of Table 4.I. This dataset, from the RST point of view, results non-consistent, having a dependency degree of 0.975.

The QRA returned 9 variables: the HHb power in the VLF band after BH (P_{VLF} post BH – HHb), the O_2Hb power in the LF band before HYP (P_{LF} pre HYP – O_2Hb), the HHb power in the LF band after HYP (P_{LF} post HYP – HHb), the coherence value in the VLF band before BH (S_{VLF} pre BH), the coherence values in the VLF and LF band before and after HYP (S_{VLF} pre HYP, S_{VLF} post HYP, S_{LF} pre HYP, S_{LF} post HYP), and the BHI_{O2Hb}. A one in the second column of Table 4.I indicates the selected features. The reduced set of features presents a dependence degree equal to the value of the initial set of variables, that is 0.975.

The third column of Table 4.I reports the results of the ANOVA analysis considering the subject pathology as independent variable and the 26 previously described parameters as dependent variables. In this study, only the features with p-values lower than 5% were deemed as mainly descriptive of the subjects' classification. This allowed removing the features with a poor correlation with the independent variably (i.e. migraine type). Three were the parameters emerged from this analysis (marked with asterisk in third column of Table 4.I): the O_2 power in the LF band post HYP (P_{LF} pre HYP – O_2Hb), the BHI_{O2Hb}, and the BHI_{HHb}. Only the last variable was recognized to be relevant also by QRA procedure. Calculating the dependence degree of this feature subset, it was obtained a value of 0.3, which is very low if compared with the maximum obtainable value.

In order to test the performance of the two FS procedures and comparing them with the initial set of variables, three FFNNs were used and the results for each network are reported in Figure 4.5.

Using the whole set of features, a correct classification rate of 100% of subjects is obtained (Figure 4.5 panel A). Using the 9 parameters highlighted with the QRA, the classification accuracy of the subjects is 97.5% (Figure 4.5 panel B). The classification performance definitely decreases when FFNN takes as input the three features selected with ANOVA analysis. In this case, the correct classification rate drops down to about 75% (Figure 4.5 panel B).

Table 4.I Results of two FS procedures. The first column contains the 26 variables used as input. In the second column, results of QRA are reported (1: feature selected). Results of one-way ANOVA analysis, in terms of p-value, are collected in the third column. The significant parameters (p-value < 5%), obtained considering as independent variable the subject pathology, are indicated with asterisk. The last rows report the number of features and the dependency degree of both subsets.

Feature	QRA	ANOVA (p-value)
P_{VLF} preBH O_2Hb	0	74.09%
P_{VLF} postBH O_2Hb	0	65.42%
P_{LF} preBH O_2Hb	0	84.20%
P_{LF} postBH O_2Hb	0	22.18%
P_{VLF} preBH HHb	0	51.45%
P_{VLF} postBH HHb	1	71.68%
P_{LF} preBH HHb	0	21.96%
P_{LF} postBH HHb	0	50.65%
P_{VLF} preHYP O_2Hb	0	17.16%
P_{VLF} postHYP O_2Hb	0	70.92%
P_{LF} preHYP O_2Hb	1	69.94%
P_{LF} postHYP O_2Hb	0	**2.94%***
P_{VLF} preHYP HHb	0	73.27%
P_{VLF} postHYP HHb	0	55.18%
P_{LF} preHYP HHb	0	94.65%
P_{LF} postHYP HHb	1	46.71%
S_{VLF} preBH	1	79.05%
S_{VLF} postBH	0	75.17%
S_{LF} preBH	0	38.53%
S_{LF} postBH	0	84.55%
S_{VLF} preHYP	1	13.66%
S_{VLF} postHYP	1	90.97%
S_{LF} preHYP	1	62.81%
S_{LF} postHYP	1	27.25%
BHI_{O2}	1	**4.93%***
BHI_{CO2}	0	**0.00%***
# of features	**9**	**3**
Dependency degree	**0.97**	**0.30**

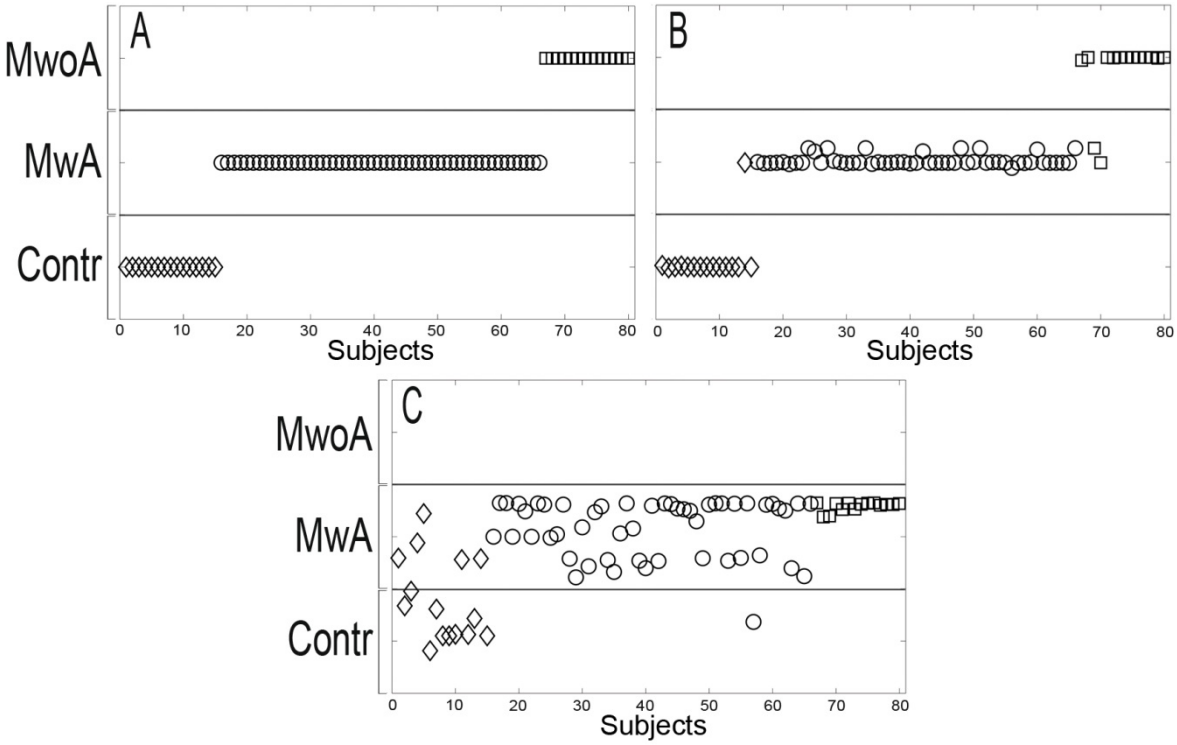

Figure 4.5 FFNN results in terms of subjects' classification. Diamonds represent healthy subject, circles are related to MwA while squares indicate MwoA. A) FFNN outputs using all features as inputs. B) FFNN outputs obtained with the 9 features selected by QRA. C) FFNN outputs related to the 3 features highlighted with ANOVA analysis.

4.4 Discussion

4.4.1 Technical Aspects

The main purpose of application was to compare the performances of a classical statistic approach (ANOVA) with a new method based on RST (QRA). Both methods were applied for identifying a minimal subset of variables able to keep the same amount of information contained in the whole set of parameters derived from NIRS signals. Such procedure allows emphasizing the attributes that are relevant to obtain a reliable classification of the subjects.

The subset performances were assessed by means of FFNN, a supervised procedure in which knowledge is acquired by the network through a learning process.

As it emerges from the results reported in Table 4.I, even if QRA returns a subset with higher cardinality than ANOVA, from the dependence degree point of view there is no loss of information in the variables selected by QRA with re-

spect to the total set of features. This is not true for the three parameters emerged from ANOVA, which leads to a reduction of the dependency degree of about 60%.

This situation is also proved by the classification results showed in Figure 4.5: QRA provides a classification rate of 97.5%, with only three subjects over 80 classified in the wrong class, while ANOVA has a classification capability of about 75%, with only 71 elements correctly classified. Moreover, in this last case it can be observed that all elements incorrectly classified are inserted into the MwA class, which is the most represented class. This is certainly due to the great imbalance in the numerosity of three classes in the dataset, which could influence the training step and consequently the classification performances. The same findings were found by Mazurowski and Habas [16].

Finally, ANOVA needs the selection of a significance threshold, which is here fixed equal to 5%. Therefore, all the features that did not explain at least 5% of the data variance with respect to the independent variable were discarded. This selection is arbitrary and unsupported by any clinical observation. On the contrary, QRA is a completely automatic procedure for FS that does not require any choice of parameters for the selection.

All these observations lead to the conclusion that, from the FS point of view, QRA gives the best results in terms of subset of features, because it is able to cope with non-linearities and it is less arbitrary. ANOVA, as all statistical methods, requires data following some strict assumptions, well defined and plentiful [17]. In fact, it is not able to correctly deal with non-linear correlation among the variables, as it happens for most real applications above all in medical field.

4.4.2 Physiological Aspects

From a physiological point of view, the results of QRA are in agreement with previously published results. It is widely accepted that migraine (and, particularly, MwA) is pathology with a vascular component involving carbon dioxide disregulation [6, 15, 24]. Among the 9 features that were considered as most important by QRA, 5 are related to the coherence between oxygenated and reduced hemoglobin in vaso-active maneuvers. For example, it is interesting to notice that MwA differed from MwoA sufferers from the coherence levels before and after hyperventilation, in both VLF and LF band. HYP is a stimulus causing a strong vasoconstriction and, therefore, a strong autoregulation is needed in order to maintain the correct proportion between the two hemoglo-

bin types in the brain tissues. Being MwA associated with impaired autoregulation [6, 7, 18], this study shows that QRA effectively found the most important variables associated to the physiological system under analysis.

4.5 Conclusions

As often happens in the analysis of complex physiological systems, many different variables must be considered in order to gain a description of the system that is as complete as possible. When the physiological systems comprises pathology (or, as in this study, two different pathologies) large feature datasets are required. In this study, the vascular pattern of subjects suffering from MwA and MwoA was analyzed by means of the time-frequency analysis of cerebral oxygenation in NIRS signals. Overall, we derived a dataset of 26 variables on which we applied two FS strategy, one based on RST while the other is a traditional statistical procedure.

QRA extracted 9 features, meaning that more than 65% of data are removed from the original set of variables against a decrease in the classification accuracy of only 2.5%. Moreover, the selected features were relevant for the pathology, since they were mostly correlated to the carbon dioxide disregulation typical of migraine with aura sufferers. ANOVA returns only three variables, which are not able to correctly identify the subject class.

This study, published in [19],represented the first and simplest application of advanced FS methods to a real set of data derived from physiological measures. It allowed highlighting the limitations of statistical approaches in dealing with real data, that often are incomplete and do not follow any specific model.

References

1. Kruit, M.C., et al., *Migraine as a risk factor for subclinical brain lesions.* JAMA, 2004. **291**(4): p. 427-34.
2. Scher, A.I., et al., *Cardiovascular risk factors and migraine: the GEM population-based study.* Neurology, 2005. **64**(4): p. 614-20.
3. Tietjen, G.E., *Migraine as a systemic vasculopathy.* Cephalalgia, 2009. **29**(9): p. 987-96.
4. Liboni, W., et al., *Why do we need NIRS in migraine?* Neurol Sci, 2007. **28 Suppl 2**: p. S222-4.

5. Nowak, A. and M. Kacinski, *Transcranial Doppler evaluation in migraineurs.* Neurol Neurochir Pol, 2009. **43**(2): p. 162-72.
6. Vernieri, F., et al., *Increased cerebral vasomotor reactivity in migraine with aura: an autoregulation disorder? A transcranial Doppler and near-infrared spectroscopy study.* Cephalalgia, 2008. **28**(7): p. 689-95.
7. Molinari, F., et al., *Time-Frequency Characterization of Cerebral Hemodynamics of Migraine Sufferers as Assessed by NIRS Signals.* EURASIP Journal on Advances in Signal Processing, 2010. **2010**.
8. Molinari, F., et al., *Relationship between oxygen supply and cerebral blood flow assessed by transcranial Doppler and near-infrared spectroscopy in healthy subjects during breath-holding.* J Neuroeng Rehabil, 2006. **3**: p. 16.
9. Silvestrini, M., et al., *Basilar and middle cerebral artery reactivity in patients with migraine.* Headache, 2004. **44**(1): p. 29-34.
10. Watanabe, Y., et al., *Monitoring cortical hemodynamic changes after sumatriptan injection during migraine attack by near-infrared spectroscopy.* Neurosci Res. **69**(1): p. 60-6.
11. Viola, S., et al., *Pathophysiology of migraine attack with prolonged aura revealed by transcranial Doppler and near infrared spectroscopy.* Neurol Sci. **31 Suppl 1**: p. S165-6.
12. Obrig, H., et al., *Spontaneous low frequency oscillations of cerebral hemodynamics and metabolism in human adults.* NeuroImage, 2000. **12**(6): p. 623-639.
13. Sliwka, U., et al., *Spontaneous oscillations in cerebral blood flow velocity give evidence of different autonomic dysfunctions in various types of headache.* Headache, 2001. **41**(2): p. 157-63.
14. *Classification and diagnostic criteria for headache disorders, cranial neuralgias and facial pain. Headache Classification Committee of the International Headache Society.* Cephalalgia, 1988. **8 Suppl 7**: p. 1-96.
15. Okada, E., et al., *Theoretical and experimental investigation of near-infrared light propagation in a model of the adult head.* Applied Optics, 1997. **36**(1): p. 21-31.
16. Mazurowski, M.A., et al., *Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance.* Neural Networks, 2008. **21**(2–3): p. 427-436.
17. Triantaphyllou, E., *Data Mining and Knowledge Discovery via Logic-Based Methods: Theory, Algorithms, and Applications*. 2010, New York, NY: Springer Publishing Company, Incorporated. 350.
18. Liboni, W., et al., *Spectral changes of near-infrared spectroscopy signals in migraineurs with aura reveal an impaired carbon dioxide-regulatory mechanism.* Neurol Sci, 2009. **30 Suppl 1**: p. S105-7.
19. Rosati, S., G. Balestra, and F. Molinari, *Feature Extraction by QuickReduct Algorithm: Assessment of Migraineurs Neurovascular Pattern.* Journal of Medical Imaging and Health Informatics, 2011. **1**(2): p. 184-192.

Chapter 5: Comparison among Different Non-Statistical FS methods

5.1 Introduction

After having highlighted the limitation of the statistic methods for FS in dealing with real data, now we aim to compare the performances of different methods for FS not based on statistics. In particular, all techniques taken into account in this study were based on RST and applied for the characterization of diabetic oxygenation patterns during ankle flexo-extensions.

Computer methods have been extensively used to aid the clinical assessment of diabetes-related pathologies, such as neuropathy [1] and sympathetic system impairment [2], retinopathy and eye fundus damage [3], central and peripheral reduced perfusion [4], and gait abnormalities [5]. Beside electromyography, also near-infrared spectroscopy (NIRS) was used to monitor muscle oxygenation in subjects suffering from lower-extremity arterial disease [6] and diabetes [7]. In fact, NIRS is a low-cost, non-invasive and portable system that is very adapted for monitoring during rest and exercise conditions.

NIRS was used to monitor the oxygen and carbon dioxide concentration changes in the muscle tibialis anterior of diabetic patients and controls during a simple exercise of ankle flexo-extension, in order to evidence if NIRS parameters could be indicative of the peripheral vascular and metabolic pattern in the subjects. NIRS signals, however, require advanced signal processing procedures, because relevant information is carried by the signals' spectrum. Being such signals non-stationary when recorded during muscle activation (*i.e.*, mus-

cle contraction), we adopted a time-frequency based analysis procedure. This analysis technique allows the extraction of several spectral features that make the results of difficult interpretation. Multivariate analysis and dimensionality reduction approaches could be useful to improve the readability of the results [8].

In this study, four FS algorithms were applied to a set of features extracted from NIRS signals recorded in diabetic subjects during ankle flexo-extension, with the aim of extracting the relevant NIRS features, which were characteristic of the subjects' peripheral vascular status.

5.2 Methodology

5.2.1 Demographics

This study involved 39 type-II diabetic subjects. Each of them performed daily physical activity for one year: 19 subjects carried out adapted physical activity (APA) (age: 66.7 ± 5.7), and 20 patients performed fit walking (FW) (age: 66.0 ± 6.2). Moreover a group of 16 healthy subjects (Contr) (age: 65.3 ± 3.9) was included in the study as control.

The study received the approval from the Institutional Review Board of Gradenigo Hospital of Turin (Italy), where all the NIRS signals were acquired. All the subjects were instructed about the purposes of the study and signed an informed consent prior to undergoing tests.

5.2.2 Data acquisition & Experimental Protocol

NIRS signals were recorded on each subject before and after the period of physical exercise by using a commercially available NIRS system (NIRO300, Hammamatsu Photonics, Japan). The recordings of NIRS signals were performed with the emitting probe placed on the left tibialis anterior muscle, approximately in correspondence of the muscle belly. The receiving photodiode was placed 4 cm apart, aligned with the emitters in the direction of the muscle fibers. In this work three NIRS signals were considered: the changes in the concentration of oxygenated hemoglobin (O_2Hb), changes in the concentration of deoxygenated hemoglobin (HHb) and the tissue oxygenation index (TOI) defined as the ratio of oxygenated to total hemoglobin.

The experimental protocol was structured as follows:

- ✓ one minute of resting;

- ✓ ankle flexo-extension for 3 minutes;
- ✓ a final resting period of 3 minutes in order to observe the recovery phase.

The total duration for the examination was about 7 minutes.

5.2.3 Feature Extraction

In Figure 5.1, the left panels show an example of NIRS signals recorded on a healthy subject performing ankle flexo-extension. Figure 5.1 panel A1 reports the O_2Hb concentration change during time, panel B1 shows the HHb signal, and in panel C1 the TOI signal is depicted. The vertical dashed lines mark the onset and offset of the flexo-extension. The hemoglobin concentration significantly varies during time: flexo-extension corresponds to a decrease in the O_2Hb and an increase in the HHb concentrations.

We performed a spectral analysis of the NIRS signals, because the spectrum of NIRS signals contains the signature of the sympathetic and parasympathetic nervous drive [9].

However, as it emerges from Figure 5.1, NIRS signals are clearly non-stationary during the periods of interest. For this reason, the Choi-Williams transform (§3.4) was used in this study. The value of σ was set equal to 0.5 for all signals, as for the analysis of several biological signals.

In the right side of Figure 5.1 a Choi-Williams representations of the three NIRS signals during the experiment for a healthy subject are reported.

All the signals were preprocessed converting them to the analytical representation with zero mean. Moreover, a high-pass Chebychev filter was used, with ripple in the stopband and cutoff frequency equal to 25 mHz, in order to remove very slow components.

All time–frequency distributions were analyzed in two specific bands, VLF (20–40 mHz) and LF (40–140 mHz), before, during and after the flexo-extension. The percentage of signal power in the two bands (referred to the total power of the signal) was calculated for each period. This procedure led to obtain the following 24 variables:

- ✓ the HHb, O_2Hb and TOI power in the VLF band, before, during and after the flexo-extension (9 variables);
- ✓ the HHb, O_2Hb and TOI power in the LF band, before, during and after the flexo-extension (9 variables);
- ✓ the SCF between O_2Hb and HHb in the VLF band, before, during and after the experiment (3 variables).

- ✓ the SCF between O_2Hb and HHb in the LF band, before, during and after the experiment (3 variables).

The list of the extracted variables is reported in the first column of Table 5.I.

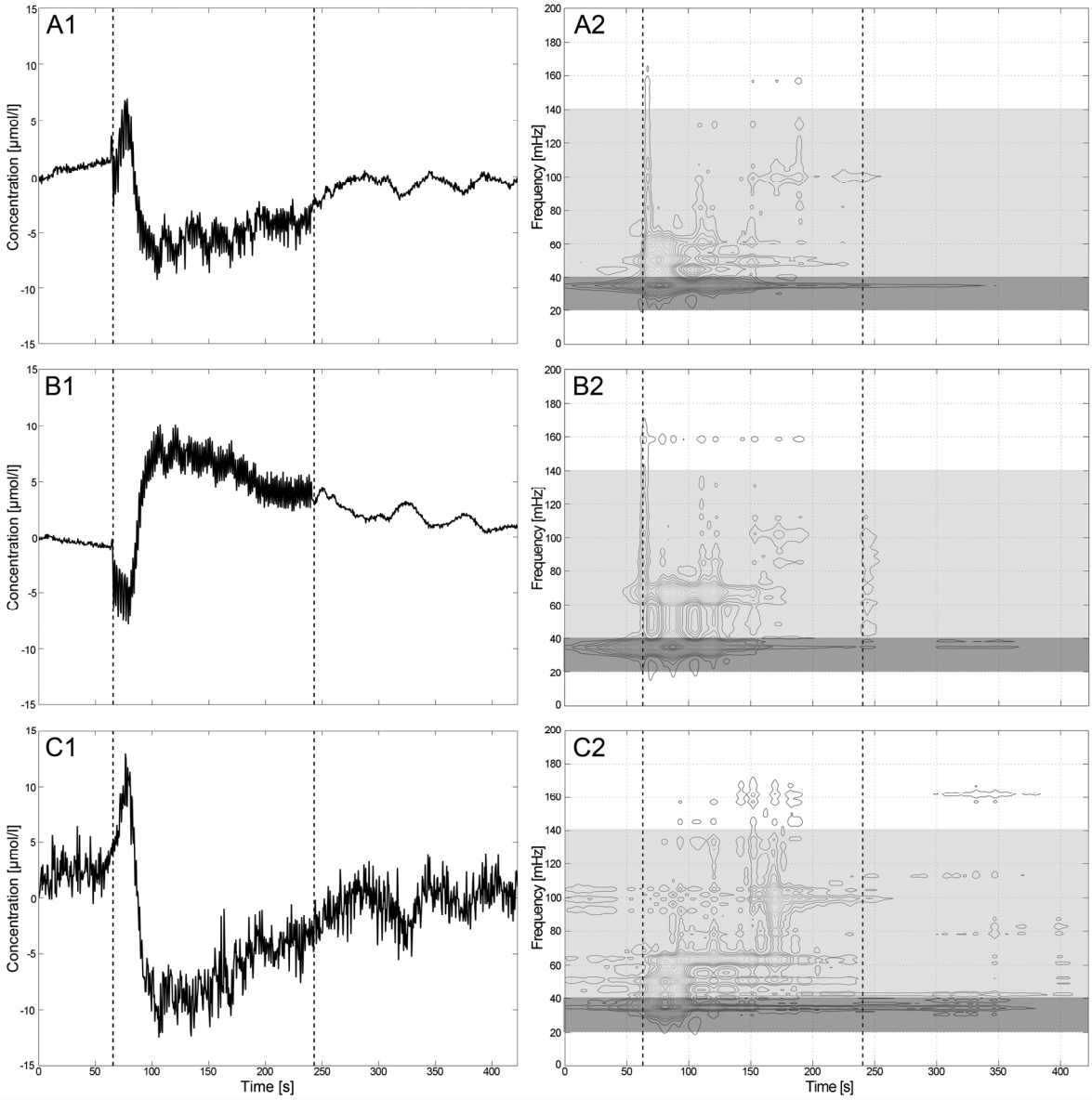

Figure 5.1 NIRS signals recorded on a healthy subject performing ankle flexo-extension. Panel A1 reports the O_2Hb concentration signal, panel B1 shows the HHb signal, and in panel C1 the TOI signal is depicted. The right panels show the 15-levels contour plot of the Choi-Williams time-frequency distribution of the signals depicted in the left side ($\sigma = 0.05$). The vertical rectangles highlight the LF band (40–140 mHz) and the VLF band (20–40 mHz). The black vertical dashed lines mark the onset and offset of the flexo-extension.

5.2.4 Feature Selection

In this study, five advanced FS procedures were performed and compared: QRA (§3.5.1), EBR (§3.5.2), GenRSAR (§3.5.3), AntRSAR (§3.5.4), and IQRA (§3.5.5). All statistical methods were discarded from this analysis because of their limitations in dealing with real and non-linearly correlated data, as proved by the first application on migraine dataset (see Chapter 4).

Regarding to GenRSAR algorithm, an initial population of 100 random individuals were generated. The population was evolved 100 times, with a probability of mutation and crossover equal to 0.4 and 0.6 respectively, as proposed in [10].To overcome the stochastic nature of this method, twenty runs of the algorithm were performed and the best solution was reported in the results, defined as the one with the higher fitness value.

Moreover, in [10] also the parameters for AntRSAR were suggested: the initial pheromone level was set to 0.5 with small random variations for each edge, α and β are set to 1 and 0.1 respectively, and the algorithm terminates after 250 iterations. Also in this case, twenty runs of the algorithm were implemented and the best subset is showed in the results as the one with the lowest cardinality, since all final subsets reach the maximum value of dependency degree.

MATLAB environment was used to implement all FS procedures.

The FS was applied to four datasets in which the rows stand for the subjects and the columns report the 24 features. The datasets were built as follows:

- ✓ diabetic patients performing APA were taken as subjects and the variables related to the signals acquired before and after the period of physical exercise represented two different classes (APA);
- ✓ diabetic patients performing FW were taken as subjects and the variables related to the signals acquired before and after the period of physical exercise represented two different classes (FW);
- ✓ features acquired before the period of exercise for control and diabetic subjects were considered as two different classes (PRE);
- ✓ features acquired after the period of exercise for control and diabetic subjects were considered as two different classes (POST);

All datasets were tested in order to detect outliers using the Wilks' method [11]. No outlier was found in any dataset.

5.2.5 Classification

The feature subsets selected by the five FS algorithms were evaluated by means of two supervised learning approaches, FFNN (§3.6.1) and SVM (§3.6.2), commonly used in the biomedical field in order to compare the performances of different FS algorithms [12].

In this study, a FFNN for each set of features and one for testing the initial variable set were built. In each network, the number of input neurons was set equal to the number of selected variables and in the output layer there was only one neuron that gives the class of the elements. Moreover, one hidden layer was inserted in the FFNN structure with a number of neurons approximately equal to 1/2 of the input neurons. As for the neuron activation functions, a logarithmic sigmoid function was used for the hidden layers and a linear function for the output layer. Back-propagation was chosen as learning algorithm and the initial values of interconnection weights were set randomly.

Similarly, six classifiers were built (one for each set of features to be evaluated and one for the whole variable set) based on SVMs using a linear kernel function.

Both FFNNs and SVMs classifiers were implemented Matlab environment.

As the aim of this study is to use the classifier in order to compare the different feature subsets, no cross-validation was performed and the whole dataset was used both for training and for testing. In this way, the FS methods are compared in the best conditions, removing the variability introduced by cross-validation. Furthermore, because of the randomness of FFNNs, each network was repeated twenty times and the results are given in terms of means value and standard deviation.

5.3 Results

In this study, five FS methods were applied to four datasets containing 24 variables characterizing diabetic subjects and listed in the first column of Table 5.I. All datasets used in this study result *consistent*, having a dependency degree equal to 1 considering the whole set of features.

The results of the five FS algorithms are reported in Table 5.I in terms of selected features, in which a 1 indicates the selected variables. Moreover, the total number of features composing each subset and their dependency degree are reported in the last rows of Table 5.I.

Table 5.I Results of the five FS procedures applied to the four datasets. First column contains the 24variables used as input for the FS. From the second to the last column are reported the results of QRA, EBR, GenRSAR, AntRSAR and IQRA applied to APA, FW, PRE, and POST datasets (1: feature selected). The selected parameters are highlighted in light grey. The last rows contain the number of features selected in each subset and the dependency degree.

	APA					**FW**					**PRE**					**POST**				
FEATURES	**QRA**	**EBR**	**GenRSAR**	**AntRSAR**	**IQRA**	**QRA**	**EBR**	**GenRSAR**	**AntRSAR**	**IQRA**	**QRA**	**EBR**	**GenRSAR**	**AntRSAR**	**IQRA**	**QRA**	**EBR**	**GenRSAR**	**AntRSAR**	**IQRA**
P_{VLF}pre O_2Hb	1	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	0	0	0	1
P_{VLF}dur O_2Hb	0	0	0	1	1	1	0	0	0	1	1	1	0	1	1	0	1	0	0	0
P_{VLF} post O_2Hb	1	1	1	0	1	0	1	1	1	1	0	0	1	0	0	1	0	0	1	1
P_{VLF} pre HHb	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0
P_{VLF}dur HHb	1	0	0	0	1	1	0	0	0	1	1	0	0	0	0	0	0	1	0	1
P_{VLF} post HHb	1	1	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0
P_{VLF} pre TOI	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	1
P_{VLF}durTOI	1	0	0	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
P_{VLF} post TOI	1	1	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
P_{LF} pre O_2Hb	0	1	1	1	0	0	1	0	1	0	0	0	0	1	0	1	1	1	1	1
P_{LF}dur O_2Hb	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0
P_{LF} post O_2Hb	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	1	0	1	0
P_{LF} pre HHb	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	1	1	1	0
P_{LF}dur HHb	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
P_{LF} post HHb	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
P_{LF} pre TOI	1	1	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
P_{LF}durTOI	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0
P_{LF} post TOI	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0
SCF_{VLF} pre	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1
SCF_{VLF}dur	0	1	1	0	0	1	1	1	0	0	1	0	0	0	0	0	0	0	1	0
SCF_{VLF}post	0	0	1	1	0	0	1	1	1	0	0	1	1	0	1	1	1	1	1	0
SCF_{LF} pre	0	0	0	0	0	1	0	0	1	1	0	1	1	1	1	0	0	0	1	0
SCF_{LF}dur	0	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0	0	0	1	0
SCF_{LF}post	0	0	0	1	0	1	0	0	1	1	1	1	1	1	0	1	0	1	1	1
# of features	8	6	5	10	6	7	6	5	11	7	8	5	5	10	7	6	7	5	9	6
Dependency degree	1	1	1	1	0.95	1	1	1	1	0.90	1	1	1	1	0.84	1	1	1	1	0.87

All FS methods return consistent subsets (with dependency degree equal to 1), except for IQRA that selects always non-consistent subsets with dependency degree higher than 0.8.The subset cardinality ranges from 5 and 11, with

GenRSAR giving always subsets with the lowest number of features and AntRSAR giving the largest subsets

To obtain information about the performances of the different FS methods from an application point of view, the subsets and the whole set of features were compared in the classification task using two different classifiers, FFNNs and SVMs.

The classification results of the six FFNNs are reported in Figure 5.2 in terms of mean value and standard deviation (on twenty runs) of the percentage of correct subjects' classification for each subset and for the whole variable set. The mean percentage of correct classification using the reduced subsets is always higher than 68%, with a maximum values of 83.5% obtained with AntRSAR subset selected from PRE. Considering all initial features, the classification rate ranges from 73.6% for FW dataset, to 85.3% obtained using PRE patients. Only AntRSAR applied to FW is able to generate performances better than all variables.

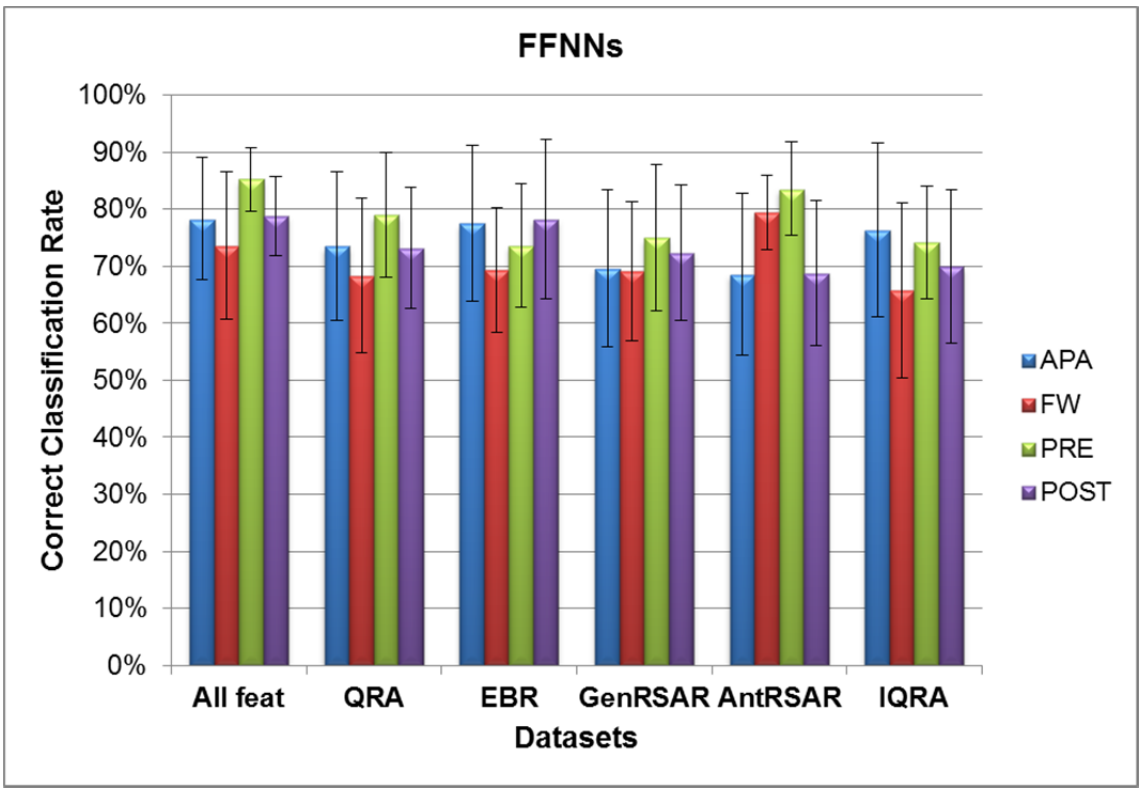

Figure 5.2 FFNN results in terms of mean value and standard deviation (on twenty runs) of the percentage of correct subjects' classification for each feature subset (QRA, EBR, GenRSAR, AntRSAR, and IQRA) and for the whole set of features (All feat.), applied to the four datasets (APA, FW, PRE, and POST).

In Figure 5.3 it is showed the percentage of correct classification obtained using the five variable subsets and all 24 features as input of six SVM classifiers. The performances obtained with SVMs are, in most of cases, slightly better than FFNN's ones. All FS methods give subsets able to correctly classify more than 69% of subjects in the five datasets, except for GenRSAR that gives a variable subset with a classification power of about 62% in the case of FW dataset. Considering the whole set of features, the percentage of correct classification is always greater than 72%, with a maximum value of 97.4% in the classification of APA subjects. In every case, no subset of variables is able to recognize a number of elements larger than the 24 parameters taken together.

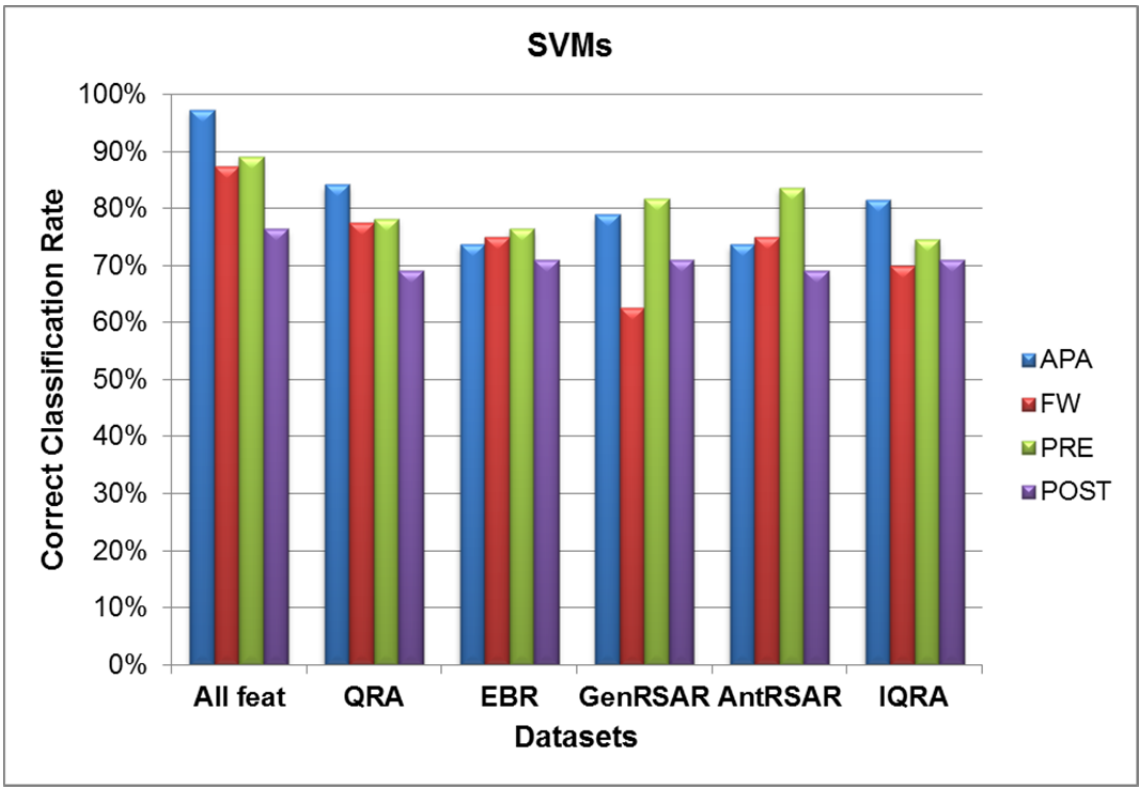

Figure 5.3 SVM results in terms of percentage of correct subjects' classification for each feature subset (QRA, EBR, GenRSAR, AntRSAR, and IQRA) and for the whole set of features (All feat.), applied to the four datasets (APA, FW, PRE, and POST).

5.4 Discussion

5.4.1 Technical Aspects

As it emerges from Table 5.I, the number of features selected from all datasets is comprised between 5 and 11, meaning that at least the 55% of the data are removed from the original datasets.

From an overall analysis of Table 5.I, it is evident the high variability in the FS results in terms of selected features, with all variables considered at least once and no parameters included in all subsets.

The dependency degree for all reducts reaches the maximum value, except for IQRA, meaning that no information was lost in most of these FS procedures from the RST point of view.

Comparing the results of the five methods for each single dataset, it emerges that most stable subsets of features are obtained for PRE and POST, in which one variable is selected by all FS tools (respectively P_{VLF} pre O_2Hb and P_{LF} pre O_2Hb). Moreover, in POST six features are never selected: P_{VLF} pre HHb, P_{VLF} dur TOI, P_{VLF} post TOI, P_{LF} dur HHb, P_{LF} pre TOI, and P_{LF} dur TOI. In particular, the power during ankle flexo-extension in the LF band for HHb and TOI are not included also in PRE subsets, meaning that these two variables are not so important for discriminating healthy subjects from pathological ones, neither before nor after the period of physical activity. For APA and FW datasets, there are no variables included in all subsets.

From these considerations, it can be concluded that the stability of the reduced subsets for the five methods tested in this work is mostly related to the sample dimension. In fact, the most stable subsets were obtained with PRE and POST datasets, made of 55 subjects (16 Contr and 39 diabetic patients). With the remaining two datasets, APA consisting of 38 samples (19 diabetic patients performing adapted physical activity before and after the period of physical exercise) and FW made of 40 samples (20 diabetic patients performing fit walking before and after the period of physical exercise), the selected subsets are very different from method to method.

Analyzing each FS method separately, it can be observed that the cardinality of the features subsets remains almost stable for each single method, even if it applied to different datasets, while results very different changing the FS algorithm. Specifically, GenRSAR always returns subsets with the lowest number of features (particularly it selects only subsets made of 5 features), whereas the reducts obtained with AntRSAR have the largest cardinality (containing 9 or 10 features).

If we relate these findings with the classification performances (Figure 5.2 and Figure 5.3), that are always above 62%, we can state that there are different combinations of the features that are able to differentiate the groups. Moreover, as the classification results obtained using all the features, both for FFNNs and SVMs, are comparable with the classification results based on the reduced sets, we can deduce that all FS tools identified discriminative parameters. This situation is coherent with the subsets' dependency degree, that reaches the highest obtainable value (1) or is near to it (0.84) in all cases. It is also clear that the classification performances are not correlated with the number of selected features.

Referring to the classification performances of the selected subsets, using FFNNs the correct classification rate is slightly lower than using SVMs. This could be because Matlab Neural Network toolbox employed in this study uses the 60% of samples randomly selected as training set, while for SVMs all samples are used for the construction of the classifier.

Considering the performances of FFNNs (Figure 5.2) for all algorithms, the lowest classification rates are mostly obtained considering FW dataset (mean classification rate of about 70% considering all subsets), while PRE allows achieving the best values (mean correct rate higher than 77%).

From the SVM point of view (Figure 5.3), the worst dataset results POST, with a mean percentage of correct classification of about 70%, while the best rates are obtained with PRE and APA, with mean values better than 78%.

5.4.2 Physiological Aspects

There is a fundamental difference between this study and some other recent researches focused on muscular NIRS [13, 14]: this signal analysis and FS procedure was based on time-frequency distributions, whereas usually NIRS signals are analyzed in the time domain. For example, De Blasi *et al.* [7] documented alterations in skeletal muscle blood flow and oxygenation in diabetic patients undergoing haemodialysis with respect to controls. Their analysis was based on the changes in the O_2Hb and HHb concentrations recorded during the haemodialysis. However, we documented a very high variability in the time course of the NIRS signals (Figure 5.1) and we did not find any significant time parameter that could discriminate diabetic patients from controls or diabetic patients undergoing different training protocols. We focused our analysis on the spectral changes of the NIRS signals, in order to understand if the pathology (diabetes) or the training (FW vs. APA) differentiated the signals' spectrum. In fact, the VLF band is directly related to the long-term metabolic adaptation and

regulation, whereas the LF band is linked to the activity of the vagal nerve and the microcirculation vassal reactivity [9]. Since the number of features that can be extracted from the time-frequency analysis is very high, we studied the potentialities of FS to extract the relevant information for the characterization of the subjects' microvascular pattern.

Analyzing Table 5.I as a whole, we can see that the FS algorithms provided solutions having a small number of variables in common. However, it is evident that the less selected parameters are related to HHb power. This is probably due to the high concentration of HHb during the test, which is not significantly different among the subjects. Studies reported that the most significant difference between diabetic and healthy muscles is the O_2Hb concentration change as detected by NIRS, rather than the HHb concentration [7].

Considering the four datasets separately, we can observe that there is less difference among the results and that each dataset is characterized by a different set of parameters. When comparing diabetic patients pre and post APA, all the most discriminative features were direct measures of tissue oxygenation (Table 5.I, first columns group) derived either from the O_2Hb or the TOI signals. Since the percentage power in the bands of the TOI signal was discriminative before (pre), during (dur), and after (post) the test contraction, we hypothesized that APA effectively improved on tissue oxygenation. This is also proved by the classification results obtained for APA dataset, that are generally higher than FW results. This means that APA produces a modification in patients' oxygenation with respect to the pre-activity period that is greater than the changes caused by FW.

If we considered the last two columns groups, which compare controls to diabetic patients before (PRE) and after (POST) training, we again observe a predominance of features computed either on the O_2Hb signal or on the TOI signal and the presence of features related to the SCF between the O_2Hb and HHb signals. This confirms the difference in the microcirculation autoregulatory pattern of the diabetics with respect to controls, both before and after physical activity. However, from the classification results it is possible to observe that the classifiers' performances are better for PRE than for POST. In this sense, as it is easier to discriminate controls from pathological subjects before the period of training, we can conclude that both kinds of exercise contribute to reduce the difference in the tissue oxygenation between diabetic and healthy people.

5.5 Conclusions

In this study, we analyzed the oxygenation pattern of the tibialis anterior muscle of a group of diabetic patients undergoing a training protocol (either FW or APA). Overall, we derived four dataset of 24 variables on which we applied five FS strategies, based on RST and not on traditional statistics.

Our findings demonstrated the utility of muscular NIRS in the assessment of the subjects' peripheral vascular pattern. In fact, we showed that by selecting the most relevant features of the NIRS signals it is possible to characterize the diabetic patients with respect to the activity they performed or against controls, with classification performance of 70% or higher. Moreover, the physical activity contributes to reduce the difference in the oxygenation patterns between pathological and healthy subjects.

This study, partially published in [15], highlighted the potentialities of advanced FS methods in dealing with real medical datasets. Moreover, it suggested the need for improved classification schemes, in order to gain a better comprehension of the diabetic muscle vascular pattern.

References

1. Acharya, U.R., et al., *Computer-based identification of type 2 diabetic subjects with and without neuropathy using dynamic planter pressure and principal component analysis.* Journal of Medical Systems, 2012. **36**(4): p. 2483-2491.
2. Desai, K., et al., *Diabetic autonomic neuropathy detection by heart-rate variability power-spectral analysis.* Journal of Mechanics in Medicine and Biology, 2012. **12**(3).
3. Rossant, F., et al., *A morphological approach for vessel segmentation in eye fundus images, with quantitative evaluation.* Journal of Medical Imaging and Health Informatics, 2011. **1**(1): p. 42-49.
4. Last, D., et al., *Global and regional effects of type 2 diabetes on brain tissue volumes and cerebral vasoreactivity.* Diabetes Care, 2007. **30**(5): p. 1193-1199.
5. Sawacha, Z., et al., *Diabetic gait and posture abnormalities: A biomechanical investigation through three dimensional gait analysis.* Clinical Biomechanics, 2009. **24**(9): p. 722-728.
6. Comerota, A.J., et al., *Tissue (muscle) oxygen saturation (StO2): A new measure of symptomatic lower-extremity arterial disease.* Journal of Vascular Surgery, 2003. **38**(4): p. 724-729.
7. De Blasi, R.A., et al., *Microcirculatory changes and skeletal muscle oxygenation measured at rest by non-infrared spectroscopy in patients with and*

without diabetes undergoing haemodialysis. Critical care (London, England), 2009. **13 Suppl 5**.
8. Molinari, F., et al., *Time-Frequency Characterization of Cerebral Hemodynamics of Migraine Sufferers as Assessed by NIRS Signals.* EURASIP Journal on Advances in Signal Processing, 2010. **2010**.
9. Obrig, H., et al., *Spontaneous low frequency oscillations of cerebral hemodynamics and metabolism in human adults.* NeuroImage, 2000. **12**(6): p. 623-639.
10. Jensen, R. and Q. Shen. *Finding Rough Set Reducts with Ant Colony Optimization*. in *Proceedings of the 2003 UK Workshop on Computational Intelligence*. 2003. Bristol.
11. Wilks, S.S., *Multivariate Statistical Outliers.* Sankhyā: The Indian Journal of Statistics, Series A (1961-2002), 1963. **25**(4): p. 407-426.
12. Sree, S.V., et al., *Use of Data Mining Techniques for Improved Detection of Breast Cancer with Biofield Diagnostic System*, in *Engineering Applications of Neural Networks*, D. Palmer-Brown, et al., Editors. 2009, Springer Berlin Heidelberg. p. 444-452.
13. Ferreira, L.F., et al., *Effects of pedal frequency on estimated muscle microvascular O2 extraction.* European Journal of Applied Physiology, 2006. **96**(5): p. 558-563.
14. Hirata, K., et al., *Effects of electrical stimulation and voluntary exercise on muscle oxygenation assessed by NIRS.* Osaka city medical journal, 2006. **52**(2): p. 67-78.
15. Rosati, S., G. Balestra, and F. Molinari, *Feature selection applied to the time-frequency representation of muscle near-infrared spectroscopy (NIRS) signals: Characterization of diabetic oxygenation patterns.* Journal of Mechanics in Medicine and Biology, 2012. **12**(4).

Chapter 6: FS Applied for Image Processing

6.1 Introduction

As it emerged from the previous applications reported in this thesis, statistics is often not enable to capture the most important information in datasets, above all if they are affected by non-linearity between features and classes. In this study we applied an automatic and non-linear FS technique for image processing. In particular, the whole KDD process was used to develop a tool for the automatic segmentation of ultrasound carotid images and intima-media thickness (IMT) evaluation.

Atherosclerosis is one of the major life threatening disease in the world [1]. The effect of atherosclerosis is the loss of elasticity of the arterial wall and the deposition, within the wall itself, of lipids and other blood-borne molecules [2]. This leads, in a range of about 5-10 years, in possible impairments to the blood circulation that could damage the principal organs (*i.e.* liver, kidneys, heart and brain).

It has been shown that the IMT of the major arteries is an important indicator of atherosclerosis [3-5]. Consequently, the correct measurement of the IMT is of crucial importance for the correct evaluation of the patients' cardiovascular risk, for the early diagnosis of atherosclerosis, and for screening.

The most widely indicator is the IMT of the carotid artery (CA), which has been used in several different multi-center studies around the world [3, 5-11]. The main benefits coming from the use of IMT in the clinical practice are related to its high repeatability, its proved validity as early marker of atherosclerosis progression and the non-invasiveness of its measurement [12].

The ultrasound scan of the arterial bed is the most widely diffused clinical examination in the field of atherosclerosis prevention and monitoring [13]. Large arteries, such as the carotids, the femoral, the brachial, and the aorta, are imaged through acoustic waves in order to visualize their wall inner composition.

The broad spread of ultrasound techniques in the clinical field is related to some advantages derived from their usage. Firstly, they are based on non-ionizing radiations that have no dangerous effects from the biological point of view. Then, the examination is safe and quick for the patient, and the equipment is relatively low-cost and thus it is largely available in many hospitals and facilities. However, ultrasound images are user-dependent and present a low signal-to-noise ratio (SNR) that compromise their easy analysis and interpretation.

6.1.1 Carotid Anatomy & Vessel Segmentation

The CAs are arteries supplying the neck and the head with oxygenate blood. Their initial segment, called *common CA*, is born from the aortic arch and rises along the neck at a depth of about 2-4 cm. At the level of the fourth cervical vertebra, it splits in two bifurcations that produce the *internal* and *external CAs* (Figure 6.1).

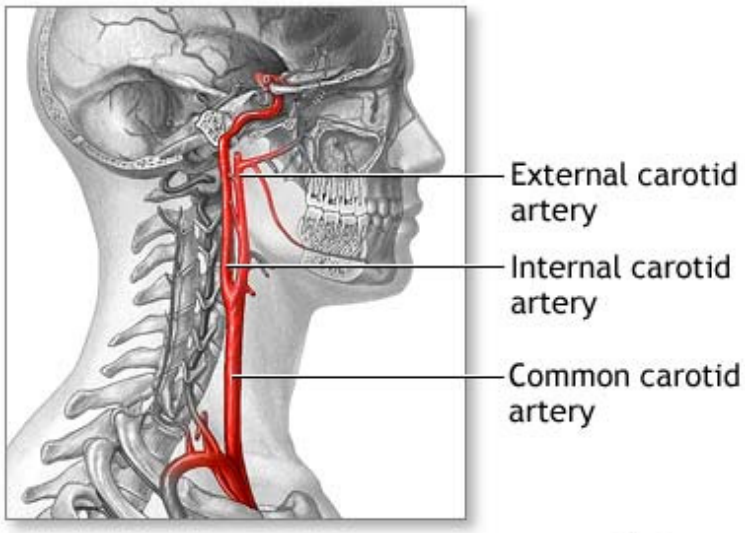

Figure 6.1 Anatomical representation of the course of the common, internal and external CAs in the supra-aortic circulation (http://www.healthcentral.com/heart-disease/13939-146.html).

Analyzing the arterial wall from the lumen towards the outside it is possible to recognize three layers:

- ✓ the *tunica intima*, that is the layer directly in contact with the blood, made of an endothelial coat followed by a layer of connective tissue and a lamina of elastic fibers;
- ✓ the *tunica media*, constituted essentially by smooth muscle cells;
- ✓ the *tunica adventitia*, including both connective tissue and elastic cells.

The segmentation of the arterial wall consists in tracing on the image the two most evident interfaces:

- ✓ The lumen-intima (LI) interface (depicted by the red line in Figure 6.2);
- ✓ The media-adventitia (MA) interface (depicted by the green line in Figure 6.2).

The interface between intima and media is not evident from ultrasound images because of the poor difference in the acoustic impedance between the two layers, and thus, it usually refers to these layers as *intima-media complex*.

The IMT is then defined as the geometrical distance (expressed in mm) between LI and MA interfaces in the common tract of the CA (Figure 6.2).

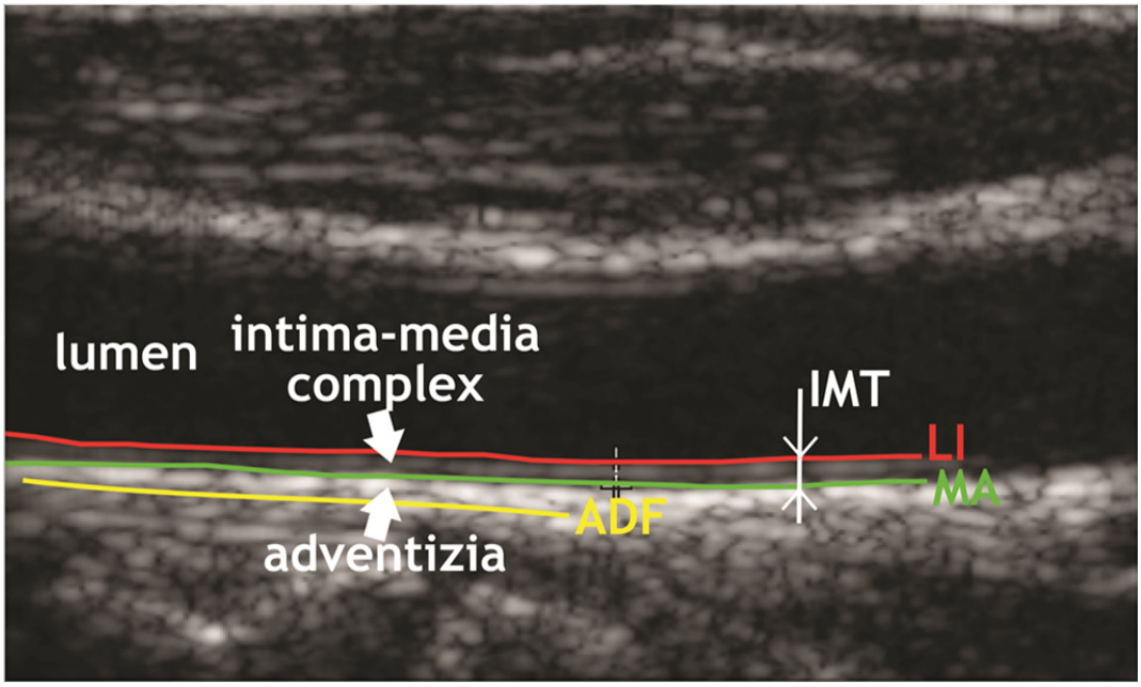

Figure 6.2 Longitudinal projection of the common tract of the CA. The red line corresponds to the lumen-intima (LI) interface, the green line marks out the media-adventitia (MA) interface, and the yellow line represent the far adventitia layer (ADF).

6.1.2 Challenges in IMT Evaluation

The clinical measurement of the carotid IMT is not a trivial task. Usually, a trained sonographer acquires a longitudinal projection of the CA and manually measures the IMT by placing two markers in correspondence of the two interfaces, LI and MA, on the image.

Manual measurements, besides being prone to errors and subjective interpretations, are time consuming and little adapt to the new quality standard nowadays requested by modern clinical best practices guidelines. For this reasons, since the pioneer work by Pignoli *et al.* [14], computer methods have been developed in order to improve the IMT measurement accuracy and to lower the inter- and intra-operator variability [15]. However, nowadays the completely automatic measurement of the carotid IMT is still an open problem.

Recently, Molinari *et al.* reviewed the most used IMT measurement techniques for the carotid wall segmentation and IMT measurement from ultrasound images [12]. The most diffused IMT measurement techniques are semi-automated, which means that the human operator interacts with the computer program in order to drive (and optimize) the IMT measurement. The interaction between expert sonographers and computer methods increase the accuracy and repeatability of IMT measurements [16], but still lacks of complete automation and brings dependence of the result on the human operator.

In the last five years, fully automated measurement techniques have foreseen a wide development. This mainly because such techniques allow for a complete independence from the user, and thus they are indicated for the processing of large datasets, typical of multi-centric and epidemiological studies. Automated techniques still underperform with respect to user-driven algorithms because of the images variability introduced by noise, vessel morphology and pathology. Moreover, it is very difficult for an automated algorithm to mimic the behavior of an expert sonographer. In fact, when human operators drive the segmentation and IMT measurement in ultrasound images, they select the optimal image region based on their skills. In this scenario, skilled operators are able to select the correct morphological region (within 1 cm from the carotid bulb) with the lower influence given by noise. It is extremely hard to reproduce the same decisional process by an automated strategy. As a consequence, usually user-independent methods segment image regions with suboptimal characteristics (*i.e.*, with artifacts, excessive noise, defocused wall layers, …) not so good and easily. To give some normative data, considering the IMT thickness of 1 mm (in presence of atherosclerosis), the average IMT measurement error for user-driven techniques can be about 0.02±0.01 mm [16], which means about 2%

of the nominal value. Despite intelligent and optimized segmentation strategies, most of the automated techniques reach performances that are about 0.03±0.10 mm [17-21], which means about 3% of the IMT value. However, it must be noted that the measurement reproducibility (*i.e.*, the standard deviation of the error) is clearly higher for automated methods (about ten times higher).

A possible solution to cope with this underperformance of automated methods is the extraction of further information from the ultrasound image. In a theoretical and optimal processing pipeline, such information should be used by the segmentation strategy to mime the human operator decision process, and thus to optimize segmentation.

The basic idea of this study is to merge the classical image segmentation techniques with intelligent systems, able to take advantage from the information content of ultrasound images at pixel level.

We started calculating a large and overabundant amount of parameters extracted from ultrasound carotid images and then applied FS in order to extract the most informative parameters for the pixels classification in one of the three carotid regions of interest: lumen, intima-media complex and adventitia. Finally, once all pixels were classified, the LI and MA interfaces were automatically traced and the IMT was calculated.

6.2 Image Database Composition

We based our work on a database consisting of 665 longitudinal B-Mode images, which is multi-institutional, multi-ethnic, multi-scanner, and multi-operator. The five image datasets were acquired independently in each institution and in different years. No standardization was made among the five sets, and each sonographer adjusted the scanner settings for each corresponding patient during acquisition. All images were discretized at 8 bits (256 gray levels) and digitally sent to a computer.

The Neurology Division of the Gradenigo Hospital of Torino (Italy) provides 200 images; the Cyprus Institute of Neurology of Nicosia (Cyprus) provides 100 images; the Department of Radiology of the University Hospital of Cagliari (Italy) provides 42 images; the Hospital de S. João do Porto (Portugal) provides 23 images; and the Department of Imaging and Interventional Radiology of the Chinese University of Hong Kong (China) provides 300 images. The complete description of the image database and the patients' demographics is reported in Table 6.I.

Our database comprised different carotid morphologies: straight and horizontal (Figure 6.3 panel A), curved (Figure 6.3 panel B), and inclined arteries (Figure 6.3 panel C). In addition, both normal carotids (i.e., with an IMT lower than 0.85 mm and no plaques) and pathologic vessels (Figure 6.3 panel D) were included. Finally, we had images with a good signal-to-noise ratio and also images with a low echogenicity of interfaces (Figure 6.3 panel E).

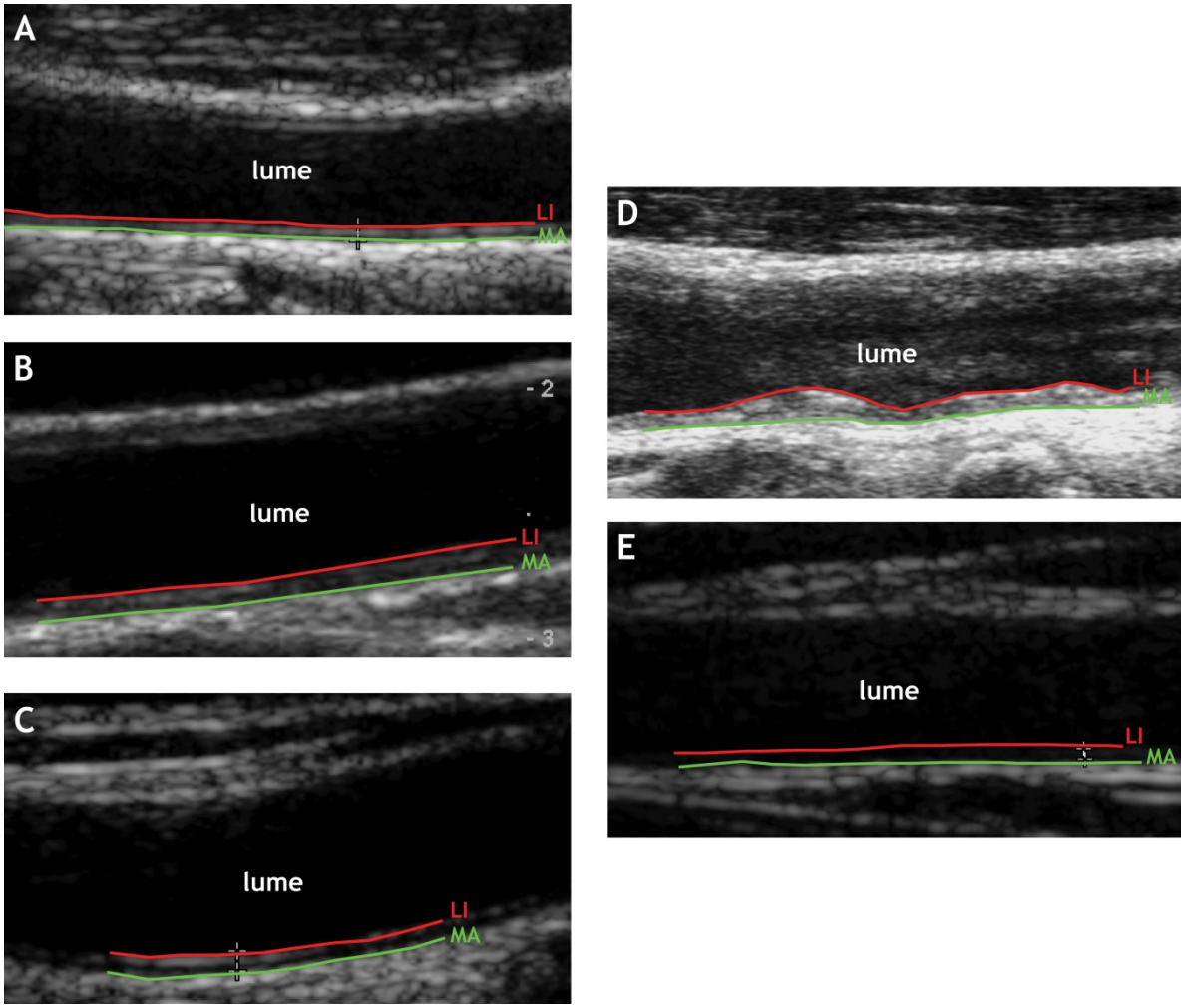

Figure 6.3 Examples of the possible carotid image typologies: straight and horizontal vessel (panel A), inclined vessels (panel B), and curved arteries (panel C), presence of plaques (panel D), and low echogenicity of interfaces (panel E). The red line corresponds to the LI interface while the green line marks out the MA interface.

Table 6.I Patient demographics and ultrasound acquisition parameters related to the whole database of images used in this study.

Institution	# Images	Conversion Factor (mm/pixel)	Scanner	# Subjects	Age (years)
Torino (2002-2009)	200	0.0625	ATLHDI5000	150	69±16 (50-83)
Nicosia (2002)	100	0.0600	ATLHDI3000	100	54±24 (25-95)
Cagliari (2009)	42	0.0789	EsaoteMyLab 70	21	68±8 (59-81)
Porto (2008)	23	0.0900	ATLHDI5000	23	-
Hong Kong (2010-2011)	300	0.0585	Siemens Antares	50	60±5 (54-67)

6.3 Feature Extraction

In this application, four classes of pixels were identified according to their physiological meaning: lumen, intima-media complex, adventitia and noisy lumen. The last class was included in order to properly manage also those images with low signal-to-noise ratio.

For each single pixel, other than its intensity, we considered as features different parameters essentially based on the intensity of the pixels around it and belonging to two categories: *statistical moments* and *texture features*. In the first class we included mean value, standard deviation, skewness and kurtosis.

The second class comprises features related to texture. Texture is the characteristic naturally used by humans when analyzing an image [22] and, in this context, texture features are a set of digital parameters based on the spatial displacement of the intensity levels in an image. They are based on the *Grey Level Co-occurrence Matrix* ($GLCM$) [23], which is calculated on an image I as:

$$GLCM_{\Delta x,\Delta y}(i,j) = \sum_{p=1}^{m}\sum_{q=1}^{n}\begin{Bmatrix} 1\ if\ I(p,q) = i\ and\ I(p+\Delta x, q+\Delta y) = j \\ 0\ otherwise \end{Bmatrix}$$

where $m \times n$ is the image size and $\delta = (\Delta x, \Delta y)$ is the displacement. This matrix simply counts the number of occurrences for which a pixel with a gray level i occurs at distance Δx and Δy from another pixel with grey level j.
Starting from the $GLCM$, the following texture features were calculated:

- ✓ *Energy:* $E = \sqrt{\sum_i \sum_j \left[SGLDM_{\Delta x,\Delta y}(i,j)\right]^2}$
- ✓ *Contrast:* $Co = \sum_i \sum_j (i-j)^2 \cdot SGLDM_{\Delta x,\Delta y}(i,j)$
- ✓ *Homogeneity:* $H = \sum_i \sum_j \frac{SGLDM_{\Delta x,\Delta y}(i,j)}{1+(i-j)^2}$
- ✓ *Entropy*: $En = -\sum_i \sum_j SGLDM_{\Delta x,\Delta y}(i,j) \cdot \log SGLDM_{\Delta x,\Delta y}(i,j)$
- ✓ *Moments:* m_1, m_3 and m_4: $m_g = \sum_i \sum_j (i-j)^g \cdot SGLDM_{\Delta x,\Delta y}(i,j)$

  where *SGLDM* represents the *Spatial Gray Level Dependence Matrix* [24] with a displacement $\delta = (0,1)$ and it is defined as:

$$SGLDM_{\Delta x,\Delta y}(i,j) = \frac{GLCM_{\Delta x,\Delta y}(i,j)}{\sum_i \sum_j GLCM_{\Delta x,\Delta y}(i,j)};$$

- ✓ *Angular Second Moment:* $ASM = \sum_{k=0}^{n-1} GLDM_\delta(k)^2$
- ✓ *Contrast:* $Con = \sum_{k=0}^{n-1} k^2 \cdot GLDM_\delta(k)$
- ✓ *Entropy:* $Ent = -\sum_{k=0}^{n-1} GLDM_\delta(k) \cdot \log GLDM_\delta(k)$
- ✓ *Mean:* $Mean = \sum_{k=0}^{n-1} k \cdot GLDM_\delta(k)$

  where $k = |i-j|$, with $k = 0,1,\dots,n-1$, n is the number of gray levels, $\delta = (0,1)$ is the displacement, and *GLDM* is the *Gray Level Difference Method* [24] calculated as:

$$GLDM_\delta(k) = \sum_i \sum_j GLCM_{\Delta x,\Delta y}(i,j);$$

- ✓ *Short Run Emphasis:* $SRE = \left. \sum_i \sum_j \frac{GLRLM_\theta(i,j)}{j^2} \right/ \sum_i \sum_j GLRLM_\theta(i,j)$

- ✓ *Long Run Emphasis:* $LRE = \sum_i \sum_j j^2 \cdot GLRLM_\theta(i,j) \Big/ \sum_i \sum_j GLRLM_\theta(i,j)$
- ✓ *Gray Level Distribution:* $GLD = \sum_i \left[\sum_j GLRLM_\theta(i,j)\right]^2 \Big/ \sum_i \sum_j GLRLM_\theta(i,j)$
- ✓ *Run Length Distribution:* $RLD = \sum_j [\sum_i GLRLM_\theta(i,j)]^2 \Big/ \sum_i \sum_j GLRLM_\theta(i,j)$
- ✓ *Run Percentages:* $RP = \sum_i \sum_j GLRLM_\theta(i,j) \Big/ A$

where A is the area of interest and *GLRLM* is the *Gray Level Run Length Matrix* [24], in which each element $GLRLM_\theta(i,j)$ represents the number of occurrences of the j adjacent elements with gray level i in direction θ equal to 0°, 45°, 90° and 135°.

Each feature was calculated on six different areas around the considered pixel, except for the intensity that was related only to the analyzed pixel. Four areas, with sizes 7x15, 15x7, 7x3, and 3x7 pixels, were centered on the selected pixel, and two non-centered areas with sizes 7x3 pixels were chosen upwards and downwards. These six regions are schematized in Figure 6.4.

The dimensions of the pixel regions were selected after a deep analysis of the context and the problem. We supposed that for pixels positioned in the inner of a region, the biggest and the smallest rectangles will have almost the same characteristics. In case of pixels near to the border between two zones, the feature values related to rectangles of different sizes will be appreciably different, and the nearest is the pixel to the interface, the highest will be the dissimilarity between small and big areas. Finally, as the interfaces we are going to find are mainly horizontal, the two non-centered rectangles were introduced to better identify the transitions between two regions in vertical direction.

With this procedure we extracted a total of 211 features for all analyzed pixels. The complete list of the image descriptors used in this work is reported in Table 6.II.

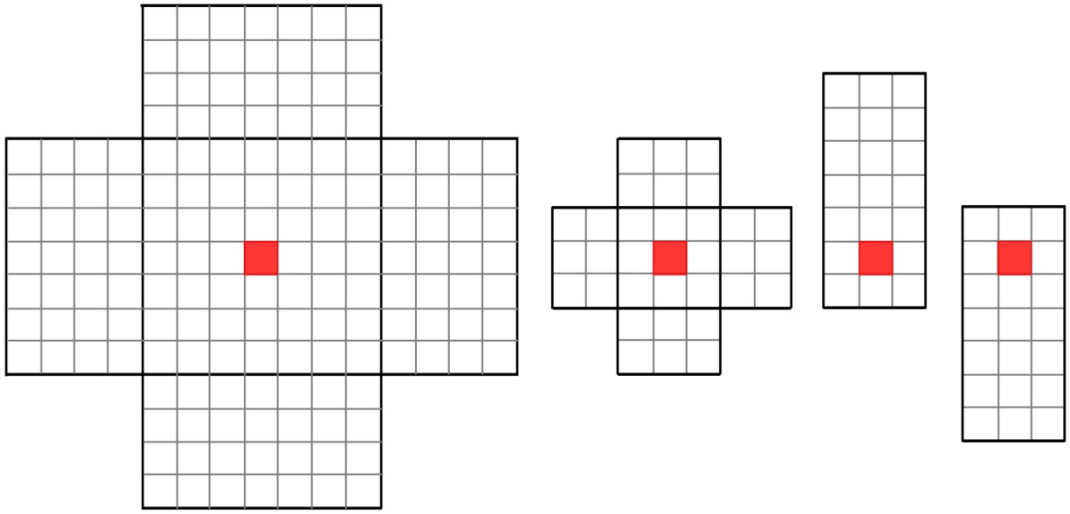

Figure 6.4 Six regions of pixels were considered around the selected pixel (colored in red) for the calculation of the features: four areas centered on the selected pixel, with sizes 7x15, 15x7, 7x3, and 3x7 pixels, and two non-centered areas with sizes 7x3 pixels upwards and downwards.

Moreover, we decided to divide the ultrasound images in small contiguous squares made of 2x2 pixels. Each area was considered as a single point with an intensity that is the mean of the intensities of the 4 pixels it includes. This choice was adopted in order to reduce the problems related to noise, that can produce substantial differences in the intensity of adjacent pixels. Additionally, in this way also the number of elements to be managed was reduced and, consequently, the computational time.

Four different datasets of pixels were built using this procedure, one used for the FS and the remaining three datasets employed as training set for the classification phase. Each dataset contained 600 pixels per class randomly chosen from 120 images randomly selected among the whole set of available images. In order to obtain pixels belonging to lumen with and without noise, 60 images were chosen with good SNR while the remaining images had low SNR. As a supervised FS method and a classification approach were chosen, each pixel was also labeled with the region it belongs to. This class was assigned considering the LI and MA profiles manually traced by an expert.

Table 6.II List of features extracted for each of 2400 pixels included in the dataset for FS. The first column contains the area on which each variable is calculated and in the second column it is reported the attribute description.

Area (pixel)	Feature Description
-	*pixel intensity*
7x15 15x7 7x3 3x7 7x3 upwards 7x3 downwards	*statistical moments*: mean value standard deviation skewness kurtosis
7x15 15x7 7x3 3x7 7x3 upwards 7x3 downwards	*SGLDM with displacement δ=(0,1):* energy contrast homogeneity entropy 1st, 3rd and 4th order moments
7x15 15x7 7x3 3x7 7x3 upwards 7x3 downwards	*GLDM with displacement δ=(0,1):* angular second moment contrast entropy mean
7x15 15x7 7x3 3x7 7x3 upwards 7x3 downwards	*GLRLM in directions θ equal to 0°, 45°, 90° and 135°:* short run emphasis long run emphasis gray level distribution run length distribution run percentage

6.4 Feature Selection

For the FS phase, the QRA (§3.5.1) was applied to the first dataset made of 2400 pixels (600 for each class) randomly extracted from 120 images and characterized by 211 features, in order to reduce the number of variables and improve the further phases of data processing.

The dataset used for FS had a value $\gamma_C(D)$ equal to 1, which means that it is consistent. Applying the QRA to the dataset, also a consistent subset of 12 parameters was obtained, listed in Table 6.III.

As it emerges from Table 6.III, at least one feature for each rectangle and for each class of image descriptor is selected, meaning that the chosen areas and features are effectively important for the description of the pixel position. Moreover, as the final subset of variables have a dependency degree equal to 1, there was no loss of information with respect to the original 211 parameters from the RST point of view.

Table 6.III Results of FS carried out by QRA applied to a dataset made of 2400 pixels with 211 features and belonging to four classes. The first column contains the considered area of pixels, and in the second column the selected features are described.

Area (pixels)	Feature Description
7x15	*GLRLM in directions θ=0°:* long run emphasis
15x7	*statistical moments*: kurtosis *GLDM with a displacement δ=(0,1):*contrast
7x3	*statistical moment*: skewness *GLRLM in directions θ= 135°:* gray level distribution
3x7	*statistical moments*: mean *GLRLM in directions θ= 0°:* run length distribution *GLRLM in directions θ= 90°:* run length distribution
7x3 upwards	*SGLDM with displacement δ=(0,1):*1storder moment *GLRLM in directions θ= 90°:* short run emphasis
7x3 downwards	*statistical moments*: standard deviation *SGLDM with displacement δ=(0,1):*1storder moment

6.5 Pixel Classification

The features subset selected using the QRA, was employed to identify the region which each pixel belongs to.

For this purpose, a classifier based on three similar FFNNs (§3.6.1) was implemented. Several different implementations of the networks were tested. The best results were obtained using an input layer made up of 12 neurons, one for each feature selected by QRA, 10 hidden layers and one output neuron. The number of neurons was progressively reduced of one element through every hidden layer. As for the neuron activation functions, we used a logarithmic sigmoid function for the hidden layers and a linear function for the output lay-

er. The schematic representation of the structure of each ANN is reported in Figure 6.5.

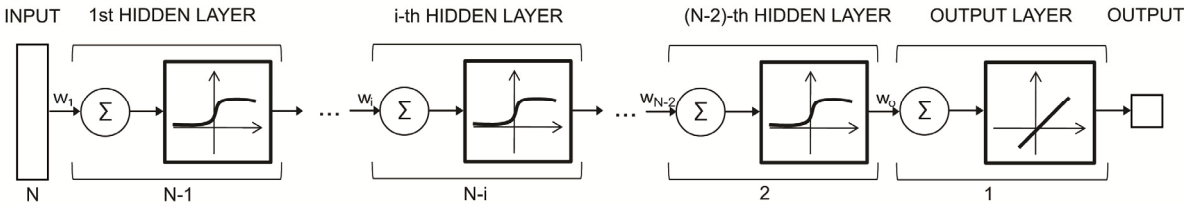

Figure 6.5 Structure of each FFNN used for pixel classification. The network consists of an input layer made up of *N* neurons (equal to the number of features selected by QRA), *N-2* hidden layers with logarithmic sigmoid activation function and with a number of neurons progressively reduced of one element so that the output layer results made up of only one neuron with a linear activation function.

Back-propagation was chosen as the learning algorithm and the mean squared error was used as performance function. The initial values of interconnection weights were set randomly. Each FFNN was trained separately using the three datasets previously constructed.

A voting system was implemented among the three FFNNs to classify the pixels of the image. First, the features associated with a single point were given as input to each network. Then, the three outcomes were used as input of a voting system returning as output the class with at least two votes. If the three network outputs were all different each other, the pixel resulted not classified. The structure of the implemented classifier is depicted in Figure 6.6.

Moreover, two different strategies were used in order to classify the pixels inside the areas of 2x2 pixels used for the feature extraction. In the first case (1st classification method – *1CM*) the output of the voting system was directly assigned to all four pixels inside the considered square (Figure 6.7 panel A). In the second approach (2nd classification method – *2CM*), each pixel was classified using a voting system implemented among the classes assigned to the four squares crossing in it (Figure 6.7 panel B).

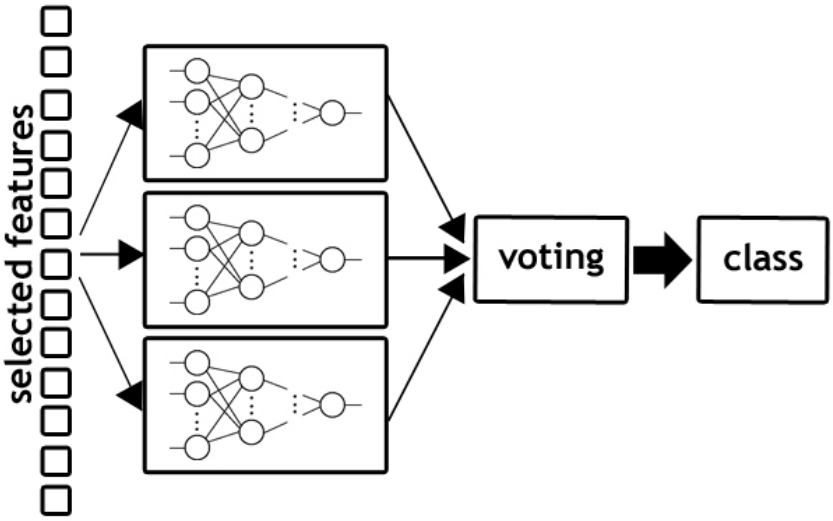

Figure 6.6 Structure of the classifier used for classifying each image pixel. The 12 features calculated for each pixel are given as input to three FFNNs. The three results are used as input of a voting system returning as outcome the class with at least two votes. If the three FFNN outputs are all different each other, the pixel results not classified.

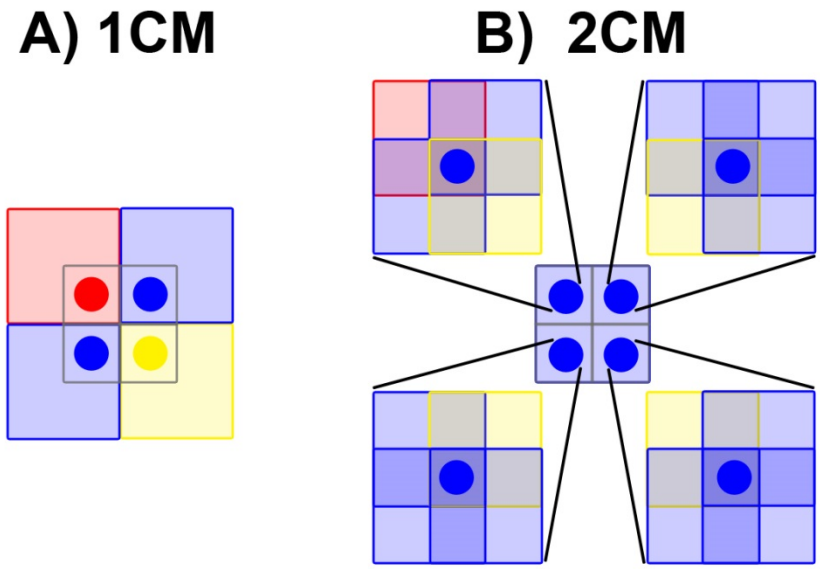

Figure 6.7 A) 1st classification method (1CM): the class assigned to the area of 2x2 pixels is automatically assigned to all pixels inside it. B) 2nd classification method (2CM): each pixel is classified using a voting system implemented among the classes assigned to the four squares crossing in it.

In order to identify the region of interest, that is the region containing the pixels to be classified, we started from the far adventitia layer (ADF) profiles automatically traced using the *Completely Automated Multiresolution Edge Snapper* (CAMES) method proposed by Molinari *et al.* [25]. Each image was then

considered column by column, moving 4 mm upwards of the ADF interface. As the entire wall thickness (intima, media and adventitia layers) is usually lower than 1.5 mm, we were sure that this region included adventitia, intima-media and lumen pixels. An example of the region of interest containing the pixel to be classified is showed in Figure 6.8.

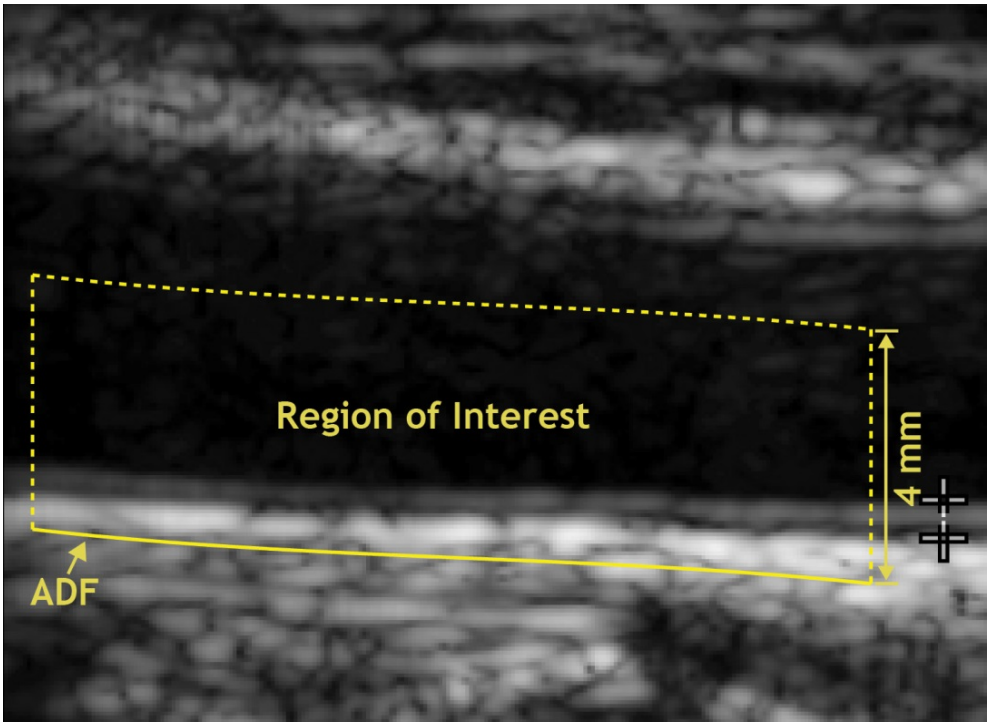

Figure 6.8 Region of interest containing the pixel to be classified. The region is identified starting from the ADF profile and moving 4 mm upwards. The considered region includes far adventitia pixels, intima-media pixels and lumen pixels.

In Figure 6.9 it is reported an example of pixel classification obtained with the first (Figure 6.9 panel A) and the second (Figure 6.9 panel B) classification method. The red points identify the lumen pixels, the blue points mark the intima-media complex pixels, the green points represent the adventitia pixels and the yellow points are not classified pixels. As it can be observed from Figure 6.9, both methods returned similar results in terms of pixel classification. Thus, we decided to conduct the two approaches for the following phase of image segmentation.

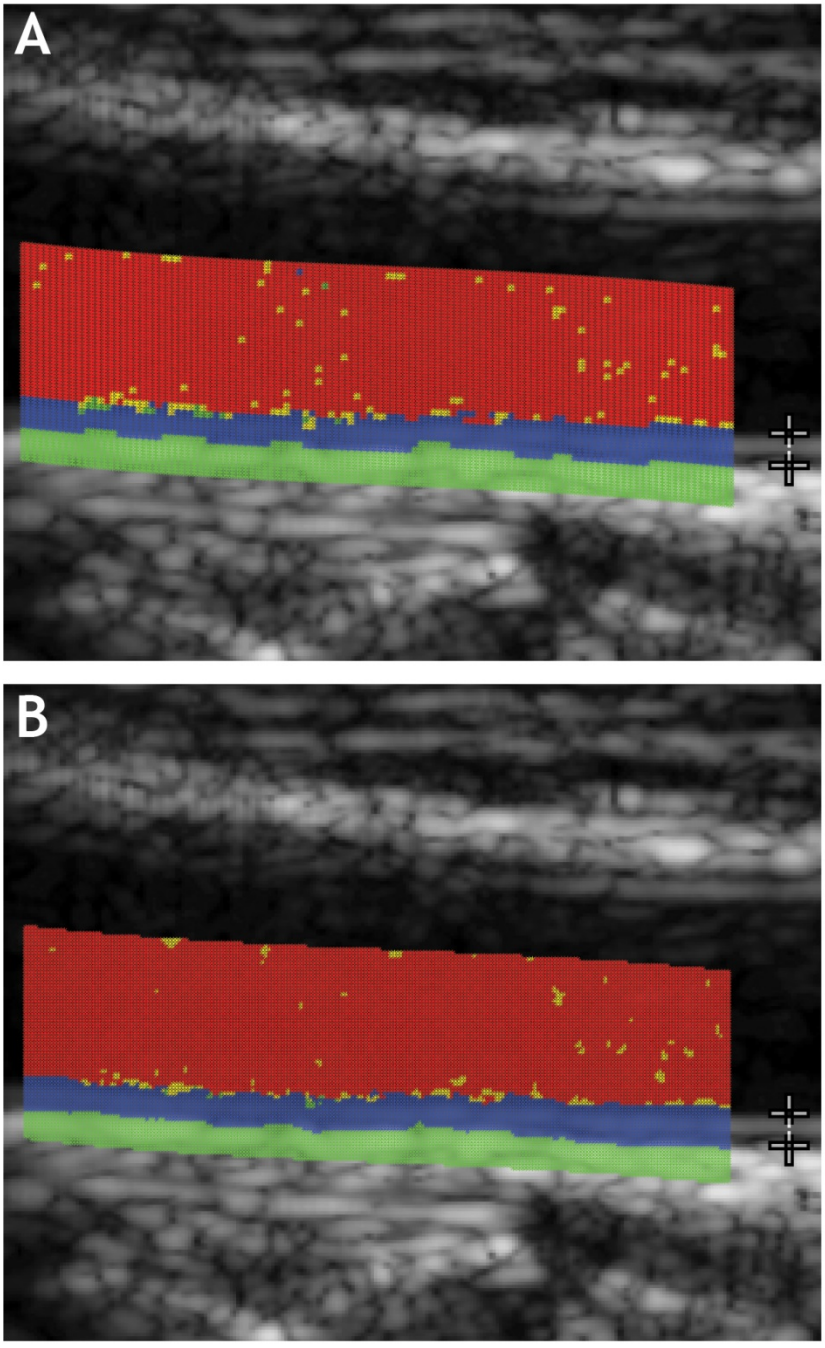

Figure 6.9 Example of pixel classification obtained with the 1CM (panel A) and the 2CM (panel B). The red points identify the lumen pixels, the blue points mark the intima-media complex pixels, the green points represent the adventitia pixels and the yellow points are not classified pixels.

6.6 IMT Evaluation

Once all pixels in the region of interest were classified, the LI and MA interfaces were identified as the sequence of the points staying between lumen and intima and between media and adventitia respectively.

For this purpose, each classified image was analyzed column by column in order to detect the transition zones among the layers. Starting from the top, subsequent overlapping 7 pixels were checked until at least 4 out of the 7 pixels belong to the intima-media complex. The pixel in the middle of that range was used as the LI pixel interface. The same algorithm was used starting from the bottom to find the MA pixel interface. The procedure used for identifying the LI and MA interfaces is illustrated in Figure 6.10.

The raw LI and MA profiles were further elaborated in order to obtain smoother interfaces that follow the anatomic carotid configuration most accurately. Firstly, all profiles were downsampled keeping only one point over ten. Then, the peaks were detected and each of them was substituted with the mean value of the two adjacent points. Finally, the new points were interpolated with a cubic spline interpolation. In Figure 6.11 it is reported an example of LI and MA profiles before (panel A) and after (panel B) this post-processing phase.

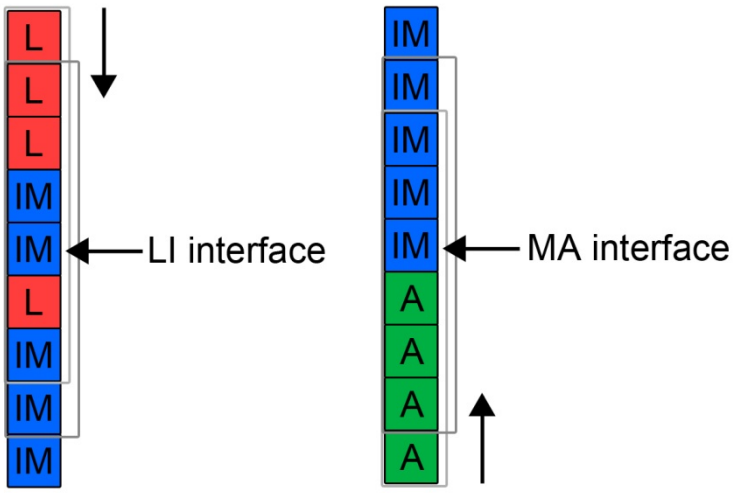

Figure 6.10 Illustration of the procedure used for identifying the LI and MA interfaces. Starting from the top of the image, subsequent overlapping 7 pixels are checked until at least 4 out of the 7 pixels belong to the intima-media complex. The pixel in the middle of that range is used as the LI pixel interface. The same algorithm id used starting from the bottom to find the MA pixel interface.

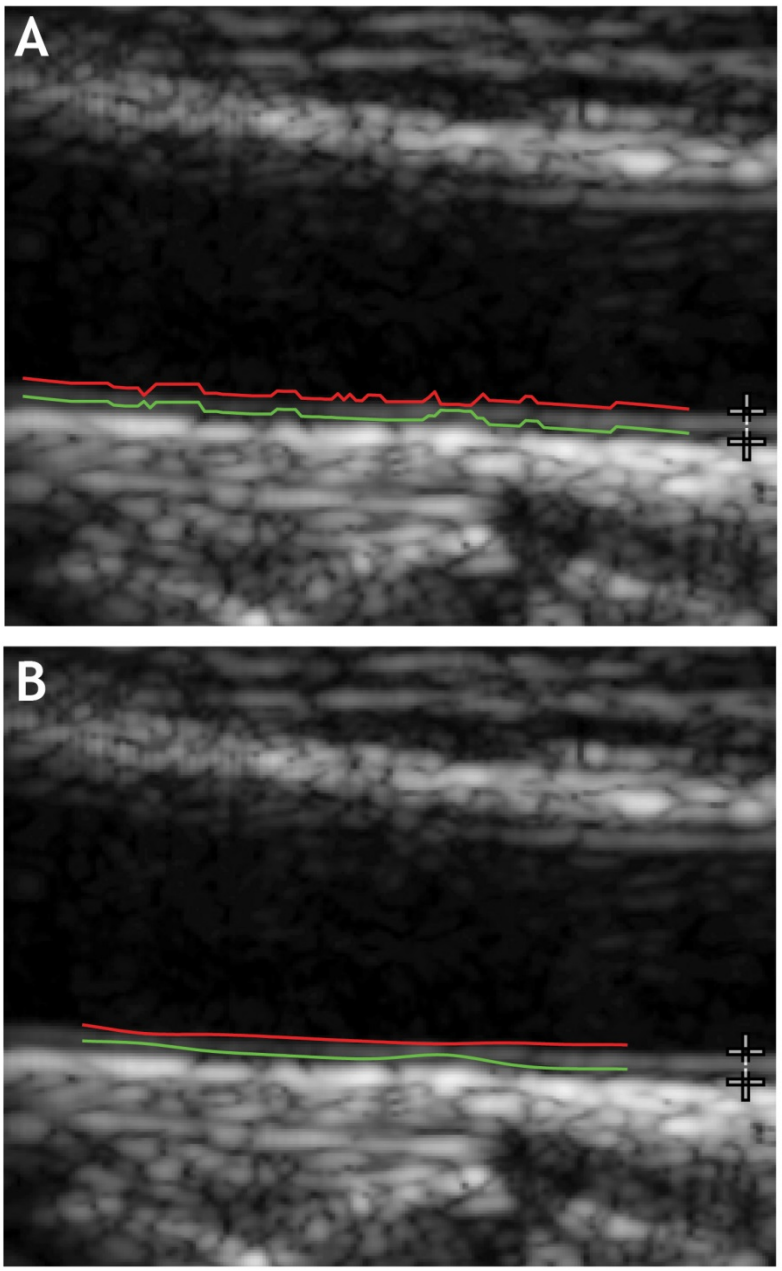

Figure 6.11 Example of LI and MA profiles before (panel A) and after (panel B) the post-processing phase consisting of downsampling, peak removal and interpolation.

Finally, the IMT was evaluated as the distance between the two profiles by using the *Polyline Distance Metric* (PDM) [26]. This measure calculates all the distances between the points in one profile and the nearest segments in the other profile. The sum of these distances is then normalized with respect to the number of points in the two lines, in order to obtain a robust metric that is independent from the number of considered vertexes.

6.7 Results

In order to assess the performances of the implemented system, we used a database of 180 images, comprising 120 images from which we selected the pixels included in the four datasets used for FS and FFNN training, 50 images randomly selected within the available database of images and not employed for the tool design and 10 images with plaques. It is important to clarify that, from the first set of 120 images only 20 pixels for image were randomly selected and included in the datasets for the system construction. As the number of classified elements for each image is about 3000 for 1CM and 12000 for 2CM, we can absolutely consider that these images will not influence the final performances of the system.

A neurosonographer and an expert vascular radiologist manually segmented all the images by tracing the LI and MA profiles. The manual segmentations were considered as Ground Truth (GT) for the system performance evaluation of the computer-generated profiles and IMT measurements. The first expert will be indicated as GT1, the second as GT2. Moreover, the performances were compared also with a traditional semi-automatic segmentation techniques based on snakes (Classical Snake – CS) proposed by Loizou *et al.* [27], that requires the manual initialization of the two profiles.

Three examples of LI (red lines) and MA (green lines) interfaces automatically traced (solid lines) and manually traced by GT1 (dashed lines) are showed in Figure 6.12, obtained using the 1CM. In Figure 6.12 panel A it is reported an image with high quality and contrast, Figure 6.12 panel B shows an image with noisy lumen, and in Figure 6.12 panel C it is showed an image with inclined vessel. For the 2CM, similar results were obtained in terms of extracted profiles.

The LI and MA profiles automatically traced with 1CM and 2CM and the semi-automatic segmentation obtained by CS were compared with GT1 and GT2. The distances between the two interfaces are showed in Table 6.IV, in terms of mean value and standard deviation. As it emerges from this table, the average LI segmentation error for our tool is about 0.21 mm for both classification methods. This value is slightly higher than the inter-reader profile distance, which is 0.13 mm. The MA segmentation error of 1CM is about 0.14 mm and for 2CM is 0.15 mm, and they result very similar to the error between the two experts that is 0.134 mm. In both cases, the distance between the automatic and the manual interfaces is lower than the one between CS and GT1 and GT2. In this case, in fact, the error is more than 0.3 mm, although CS is a semi-automatic technique.

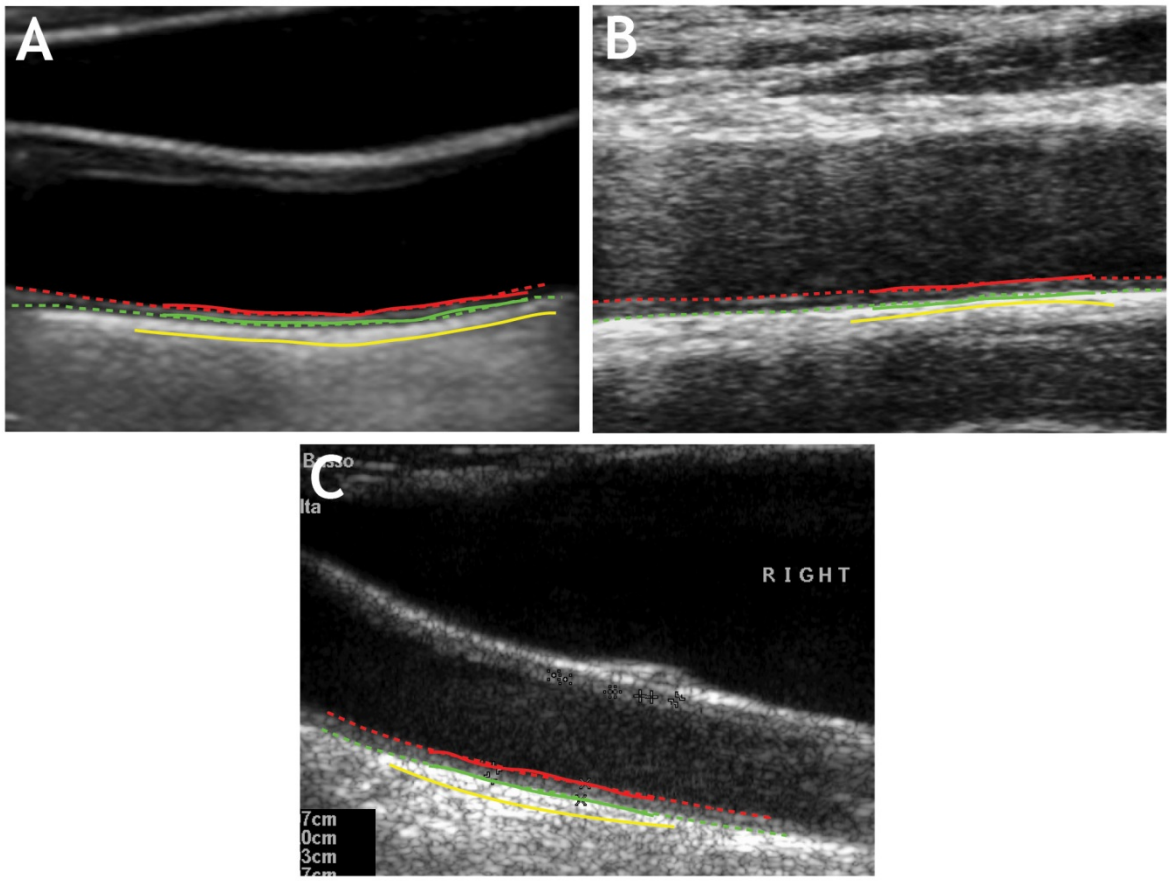

Figure 6.12 Three examples of LI (red lines) and MA (green lines) interfaces automatically traced using 1CM (solid lines) and manually traced by GT1 (dashed lines). In panel A it is reported an image with high quality and contrast, panel B shows an image with noisy lumen, and in panel C an image with inclined vessel is showed. In yellow it is traced the far adventitia boundary.

Table 6.IV Performance evaluation comparing the LI and MA interfaces obtained by the two classification methods (1CM and 2CM) and the semi-automatic classical snakes (CS) with the two sets of human-traced profiles (GT1 and GT2), in terms of distance between the profiles measured with the PDM (metric mean value ± standard deviation).

		vs. GT1	vs. GT2
GT1	**LI**	-	(0.130±0.060) mm
	MA	-	(0.134±0.100) mm
CS	**LI**	(0.343±0.214) mm	(0.321±0.207) mm
	MA	(0.335±0.264) mm	(0.323±0.259) mm
1CM	**LI**	(0.214±0.111) mm	(0.216±0.118) mm
	MA	(0.136±0.081) mm	(0.143±0.092) mm
2CM	**LI**	(0.214±0.117) mm	(0.211±0.121) mm
	MA	(0.142±0.083) mm	(0.154±0.095) mm

Table 6.V reports the errors and their absolute values, comparing the IMT obtained with 1CM, 2CM and CS with GT1 and GT2. Analyzing the IMT errors, it can be observed that the IMT measurement bias of the automatic procedure results slightly higher than the one between the two experts, especially for 2CM. Moreover, it can be observed that the automatic profiles tend to underestimate the IMT values in all cases. The IMT error absolute value obtained with our method is 0.18 mm and 0.17 mm with respect to GT1 and GT2 respectively. For CS this value increases to 0.2 mm. Moreover, also the standard deviation of the absolute error for 1CM and 2CM is significantly lower than CS in the comparison with GT1, as confirmed by the Fisher F-test ($p<1\%$). For the other comparisons, the variability is about the same and is not statistically different from the inter-operator one.

Table 6.V Performance evaluation comparing the IMT values obtained by the two classification methods (1CM and 2CM) and the semi-automatic classical snakes (CS) with the two sets of human-traced profiles (GT1 and GT2) in terms of bias and absolute error (mean value ± standard deviation).

		vs. GT1	vs. GT2
GT1	**Error**	-	(0.006±0.203) mm
	|Error|	-	(0.132±0.155) mm
CS	**Error**	(0.002±0.290) mm	(0.008±0.260) mm
	|Error|	(0.206±0.203) mm	(0.194± 0.172) mm
1CM	**Error**	(-0.110±0.222) mm	(-0.104±0.217) mm
	|Error|	(0.183±0.167) mm	(0.173±0.157) mm
2CM	**Error**	(-0.052±0.227) mm	(-0.047±0.228) mm
	|Error|	(0.177±0.151) mm	(0.173±0.156) mm

Figure 6.13 shows the histograms of the error distributions in the IMT evaluation, for the whole set of images. The bias between the segmentation tools and GT1 is reported in Figure 6.13 panels A (for 1CM) and C (for 2CM) while in Figure 6.13 panels B and D it is depicted the error of 1CM and 2CM with respect to GT2. Moreover, the normal distribution curves (dashed lines) are reported for each comparison.

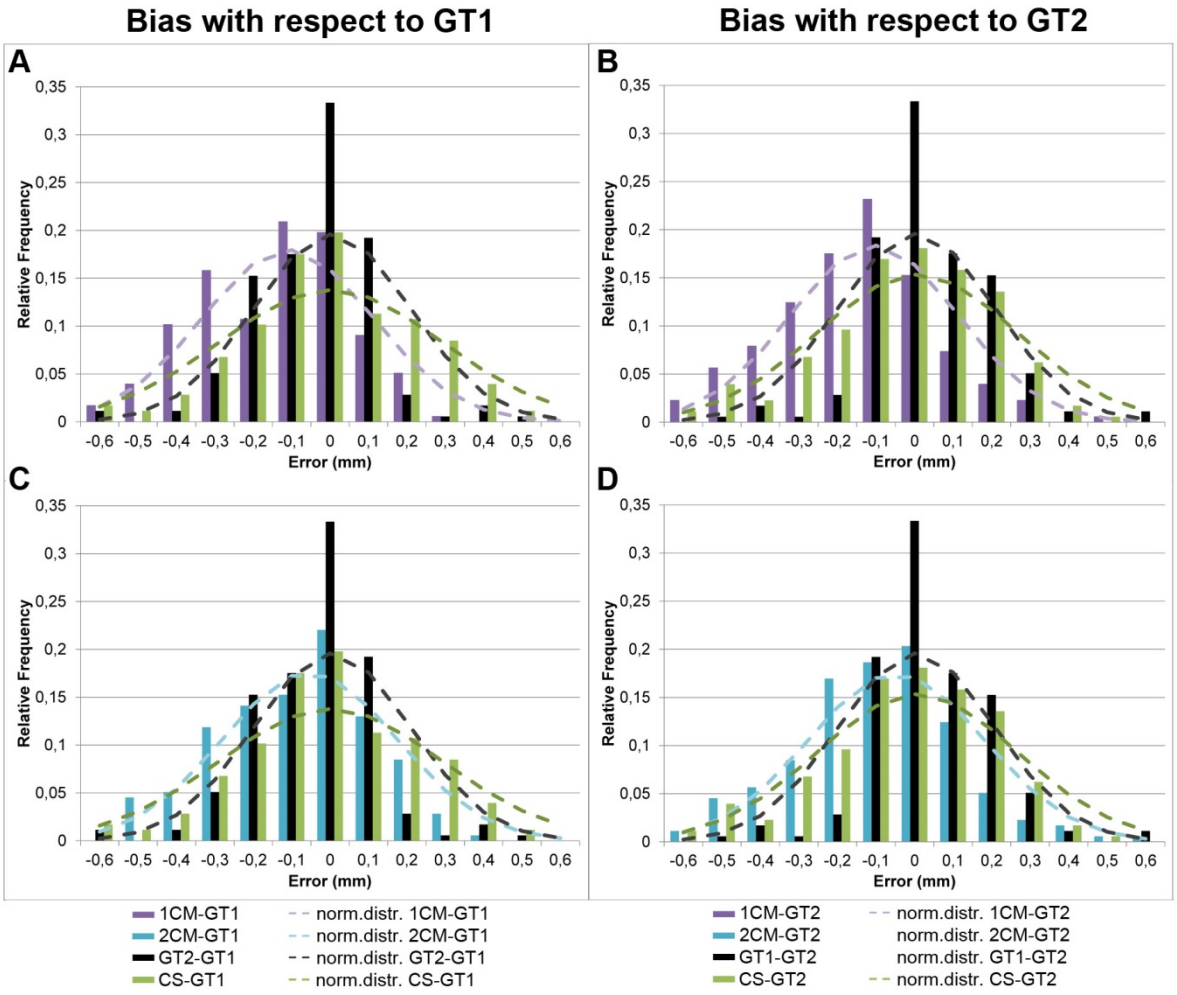

Figure 6.13 Histograms of the error distribution in the IMT evaluation for the set of 180 images. The bias between 1CM and the experts is reported in panels A (with respect to GT1) and B (with respect to GT2). In panels C and D it is reported the error between 2CM GT1 and GT2, respectively. Moreover, the distributions related to CS and the two human experts are reported as comparison. The dashed lines represent the normal distribution curves for each comparison.

Analyzing these graphs, it can be observed that the distributions obtained with our tool (violet and light blue bars) are very similar to the one related to CS (green bars), especially for 2CM (Figure 6.13 panels C and D). Moreover, the width of the normal distributions for 1CM and 2CM is lower than CS and similar to GT1vsGT2 (black bars). Finally, while 1CM presents a peak for 0.1 mm (Figure 6.13 panels A and B), 2CM has a peak in 0 mm with respect to GT1 (Figure 6.13 panel C), and GT2 (Figure 6.13 panel D).

In order to assess the relationship between our tool and the manual segmentation, the Bland–Altman plots [28] were used. This kind of graphic representations, commonly used in the clinical field for evaluating the agreement between two measurement methods, reports the difference between two sets of

measures against their mean values. This visualization allows observing any possible relationship between errors and true values (expressed in terms of mean values between the two measures). Moreover, in the Bland–Altman plot, three reference lines are traced. The first line corresponds to the mean difference between the two sets of values and it is useful in order to highlight possible biases between methods, identifying which measure tends to overestimate or underestimate the other. The other two lines identify a region comprised between 1.96 standard deviations of the differences above and below the mean value. This information allows verifying the congruence between the two methods: if the differences are normally distributed, 95% of the points will be within this interval and the metrics can be considered as congruent.

In Figure 6.14 the Bland–Altman plots are showed for the comparison between the IMT automatically evaluated using the two classification methods and the manual ones. Figure 6.14 panels A and B show the results obtained using 1CM compared with GT1 and GT2 respectively; Figure 6.14 panels C and D report the results for 2CM compared with GT1 and GT2 respectively; and in Figure 6.14 panels E and F the semi-automatic CS is compared with the two human operators. These plots show a good congruence between the outcomes of the automatic tool and the manual segmentation. In fact, it can be observe that almost all points are within the dashed lines, considering the comparison of 1CM and 2CM with both GT1 and GT2 (Figure 6.14 panels A, B, C and D). Moreover, it is evident that the IMT measurements by CS are biased when the actual thickness of the artery increases. In fact, the cloud of dots in Figure 6.14 panels E and F is decreases below the dashed bottom line, thus showing that CS underestimates IMT for carotid thickness higher than 1 mm. Overall, the CS measurements of IMT are higher than the ground truth for carotid thickness lower than 0.7 mm, and they are lower for thicker arteries. Conversely, the performance of our classification method is not biased (Figure 6.14 panels A, B, C, and D) for higher values of IMT.

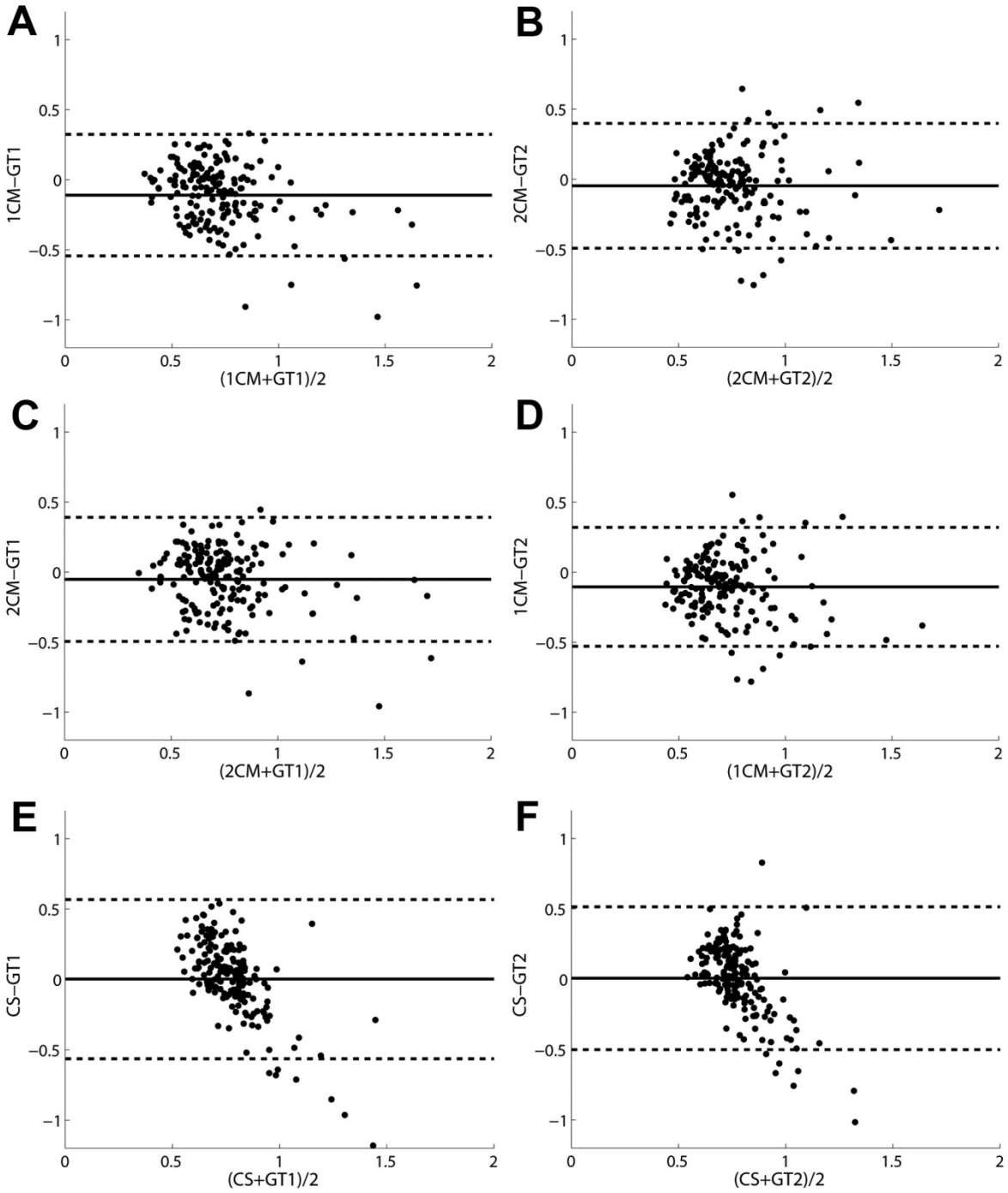

Figure 6.14 Bland–Altman plots for the two classification methods compared with the two sets of manually-traced interfaces: panels A and B show the results comparing 1CM with GT1 and GT2; panels C and D report the results for the 2CM compared with GT1 and GT2; and in panels E and F the semi-automatic CS is compared with the two human operators. The solid lines represent the mean difference between the two sets of measurements while the dashed lines are placed in correspondence of mean ± 1.96*standard deviation.

6.8 Sensitivity Analysis

With respect to the phase of feature extraction, we tested if, slightly modifying the dimensions of the pixel areas considered for the features calculation, better results in terms of IMT estimation can be obtained. Only this phase was tested in the whole process because it was based on subjective choices and not on objective evidences.

Therefore, the two biggest rectangles, with dimensions 7x15 and 15x7, were substituted with smaller ones with dimensions 7x13 and 13x7, 7x11 and 11x7, 7x9 and 9x7. Exclusively reductions in the dimensions were taken into account, because 15 pixels reach almost the thickness of the intima-media complex, that is our target. During each variation, the remaining four areas (7x3, 3x7, 7x3 upwards, 7x3 downwards) were kept unchanged.

For the phase of FS, the QRA was applied to the fist pixel dataset made of 2400 elements (600 pixels for each class) and characterized by 211 features. The results of the FS for each variation of the two biggest pixel regions (*var1* considers areas of 7x13 and 13x7 pixels, *var2* considers areas of 7x11 and 11x7 pixels, and *var3* considers areas of 7x9 and 9x7 pixels.) with respect to the initial conditions (7x15 and 15x7) are reported in Table 6.VI, in terms of selected features, subset cardinality and dependence degree.

Applying the QRA for reducing the number of features, all consistent subsets (with dependency degree equal to 1) are obtained for each variation (last row of Table 6.VI). Moreover, all subsets contain 12 features, except for the one returned for the second variation that is made of 14 variables. Only two variables are in common among the four subsets: the mean value of the intensity in the 3x7 pixels area and the first order moment calculated from the SGLDM with displacement $\delta=(0,1)$ in the region of 7x3 pixels upwards.

After the FS, the phases of pixel classification and IMT evaluation were performed on a set of 50 images not employed for the system construction. We used each new feature subset as input for the classifiers and automatically segmented the images, similarly to the procedure described above for the initial set of rectangles. The two classification approaches were again implemented.

The results, in terms of differences in the LI and MA estimation and bias in the IMT measure between the automatic tool (with both 1CM and 2CM) and the two experts, are reported in Figure 6.15 as mean value and standard deviation. Figure 6.15 panels A and B compare 1CM with GT1 and GT2 respectively, and in panels C and D the comparisons between 2CM and GT1 and GT2 are showed.

Table 6.VI Results of the FS carried out by QRA applied to a dataset made of 2400 pixels characterized by 211 features. Three variations were performed on the two biggest areas with respect to the initial conditions (7x15 and 15x7): *var1* considers areas of 7x13 and 13x7 pixels, *var2* considers areas of 7x11 and 11x7 pixels, and *var3* considers areas of 7x9 and 9x7 pixels. The first column contains the considered areas, the second column reports a description of the selected features, and in the last four columns the results of FS in terms of selected features (1) are listed. The last rows contain the number of features selected for each variation and the dependency degree of each subset.

Area (pixel)	Feature Description	Initial	Var1	Var2	Var3
7x15 - 7x13	*GLDM with a displacement δ=(0,1):* contrast				1
7x11 - 7x9	*GLRLM in directions θ equal to 0°:* LRE	1		1	
	statistical moments: standard deviation		1	1	
	statistical moments: skewness		1	1	
	statistical moments: kurtosis	1			
15x7 - 13x7	*SGLDM with displacement δ=(0,1):* 4th ord. mom.			1	
11x7 - 9x7	*GLDM with a displacement δ=(0,1):* contrast	1	1	1	
	GLRLM in directions θ equal to 0°: RP			1	
	GLRLM in directions θ equal to 90°: SRE				1
	GLRLM in directions θ equal to 90°: LRE			1	
7x3	*statistical moments*: skewness	1		1	
	GLRLM in directions θ equal to 135°: GLD	1			
	statistical moments: mean value	1	1	1	1
	statistical moments: skewness				1
3x7	*SGLDM with displacement δ=(0,1):* 1st ord. mom.		1		
	GLRLM in directions θ equal to 0°: RLD	1	1		1
	GLRLM in directions θ equal to 90°: RLD	1			
	statistical moments: skewness				1
	SGLDM with displacement δ=(0,1): 1st ord. mom.	1			1
	GLDM with a displacement δ=(0,1): contrast		1	1	
7x3 upwards	*GLRLM in directions θ equal to 0°:* GLD		1	1	
	GLRLM in directions θ equal to 0°: RLD			1	
	GLRLM in directions θ equal to 45°: SRE				1
	GLRLM in directions θ equal to 90°: SRE	1	1		
	statistical moments: standard deviation	1	1		1
7x3 down-wards	*SGLDM with displacement δ=(0,1):* 1st ord. mom.	1	1	1	1
	GLRLM in directions θ equal to 0°: GLD				1
	GLRLM in directions θ equal to 0°: RLD		1	1	1
# of selected features		12	12	14	12
Dependency degree		1	1	1	1

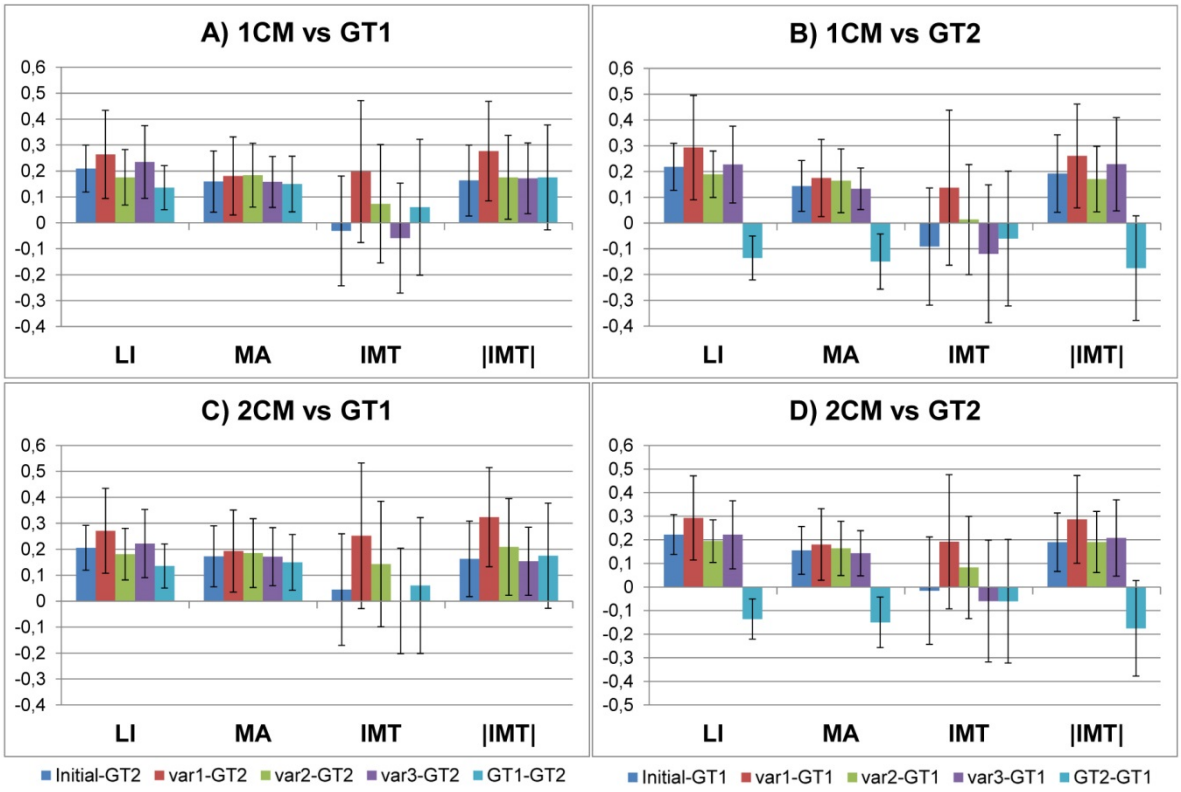

Figure 6.15 Performance comparison of the three variations in the area dimensions (var1: 7x13 and 13x7; var2: 7x11 and 11x7; var3: 7x9 and 9x7) and the initial conditions (7x15 and 15x7) with respect to GT1 and GT2. Moreover, the inter-operator variability (GT1-GT2) is reported as comparison. The results are reported in terms of differences in the LI and MA estimation and bias in the IMT measure, as mean value and standard deviation. Panels A and B compare 1CM with GT1 and GT2, and in panels C and D the comparisons between 2CM and GT1 and GT2 are showed.

Analyzing these graphs it emerges that the worst performances are obtained for *var1* (red bars), that considers areas of 7x13 and 13x7 pixels, in terms of mean value and standard deviation of all parameters, that result the highest for all comparisons. Observing the bars of the mean error in the IMT estimation and its absolute value, the best results are achieved with the initial conditions (blue bars), with rectangle dimensions of 7x15 and 15x7, confirming our initial choice about the pixel areas. This situation, in fact, gives the lowest mean IMT error in the comparison both with GT2 using the 2CM and with GT1 using the 1CM. As for the standard deviation, using the initial couple of rectangles it value is similar or in most cases lower than the dispersion between the two experts. Moreover, the mean IMT error between the 2CM and the manual profiles is lower than the averaged distance between GT1 and GT2.

6.9 Discussion

In this study, we describe the development of a new tool for ultrasound image segmentation. This technique, based on the application of feature selection and computational intelligent systems, is able to recognize the interface between LI and MA on the carotid wall, without any interaction with the user and any tuning of parameters. Nowadays, no other segmentation methodologies based on a similar approach are proposed in literature.

We proposed two versions of this tool, using different approaches for the pixel classification (1CM and 2CM). Both methods were validated on a multi-institutional, multi-ethnic, multi-scanner, and multi-operator database made of 180 images. The results were compared with one of the most used techniques for image segmentation based on snakes (CS) and two sets of manually traced profiles taken as ground truth.

As it emerges from Table 6.IV and Table 6.V, our segmentation tool based on pixel classification allows obtaining better performances than CS, although it is a semi-automatic technique, in terms of both distance between the profiles (Table 6.IV) and absolute error in the IMT evaluation (Table 6.V). However, both 1CM and 2CM tend to underestimate the IMT value, even if the second approach has a lower bias.

As for the standard deviation of the pixel classification tool, it is comparable with the inter-operators variability and definitely lower than the CS one, which is confirmed by the histograms in Figure 6.13. This means that the segmentation reproducibility, indicating the variability of the tool in returning measurement of a given quantity, is better for our technique.

The Bland-Altman plots reported in Figure 6.14 show that no bias is present between our tool outcomes and the IMT values. This is very important from the clinical point of view because it is necessary to correctly estimate the IMT both for low risk subjects (i.e. with IMT lower than 0.85 mm) and for pathological subjects (i.e. with high IMT value). On the other hand, CS tends to underestimate the IMT value with the increase of the wall thickness. This problem is crucial when snakes are used to assess the subject's cardiovascular disease (CVD) risk. In fact, advanced atherosclerosis leads to higher IMT values, which could be underestimated by a traditional snake-based algorithm (as in Figure 6.14 panels E and F). This might lead to a systematical underestimation of the CVD risk.

In conclusion, the segmentation obtained with the pixel classification tool proposed in this study, although does not produce the same performance of the

manual segmentation, gives better results than the semi-automatic methodology in terms of correctness in detecting the LI and MA interfaces and of performance reproducibility. Moreover, our novel automated system does not systematically underestimate IMT in case of advanced atherosclerosis.

With regard to the sensitivity analysis, from the FS performed by QRA we obtained all subsets with dependency degree equal to 1 (last row of Table 6.VI), meaning that there was no loss of information with respect to the original 211 parameters from the RST point of view. If we relate the fact that only two features were in common to the four subsets with the segmentation performances (Figure 6.15), we can state that there are different subsets of the features that are able to correctly identify the groups. Moreover, the classification performance is not influenced by the subset cardinality, since the most numerous subset (that is the one obtained with *var2* and containing 14 features), gives worse results than *initial* and *var3* subsets, both made of 12 features. However, from this analysis it is not only confirmed our initial selection of the rectangle dimensions for the feature extraction, but also 2CM emerges as best classification approach.

This last consideration, together with the fact that 2CM has lower bias in the IMT error with respect to 1CM (Table 6.V), leads to consider the second classification method as better than the first approach.

6.10 Conclusions

This study describes the development of an automatic tool for the segmentation of ultrasound carotid images, in order to obtain a correct estimation of the IMT. This parameter is very important from the clinical point of view because it is an indicator for cardiovascular risk.

Our method based on the pixel classification (*PxCl*) combines non-linear techniques for FS and intelligent tools for classification. In particular, the FS was implemented by QRA, that it a completely automatic algorithm able to manage incomplete knowledge. For the learning phase, the FFNNs were chosen because of their capability to extract and learn complex information from data.

The main advantages of the described tool are essentially related to the complete automation of the entire process, from the choice of the features to the LI and MA profile individuation and IMT estimation. Moreover, the implemented system works with images showing different vessel morphologies and different levels of noise, thanks to the learning abilities of the FFNNs used for the classification. In all cases, no modification or adjustment of the parameters is required.

The proposed tool showed very good performances in the interface detection and IMT evaluation, widely better than the standard semi-automatic techniques for image segmentation. Also from the clinical point of view, it has been proved that the IMT estimated with this automatic tool has the same accuracy both for normal carotid wall (with IMT lower than 0.85 mm) and in pathological conditions. This means that it can be applied for evaluating the patient risk level and for screening purposes.

Finally, as it is not yet able to reach the same outcomes than the manual segmentation, further improvements are still possible in the LI and MA identification. For example, the profiles automatically extracted with the second classification method could be used for the initialization of other algorithms, such as *active parametric contours (*or *snakes*) and evolved for improving the final performances.

This study was partially published in [29, 30].

References

1. Organization, W.H. *Cardiovascular disease*. Available from: http://www.who.int/cardiovascular_diseases/en/.
2. Badimon, J.J., B. Ibanez, and G. Cimmino, *Genesis and dynamics of atherosclerotic lesions: implications for early detection.* Cerebrovasc Dis, 2009. **27 Suppl 1**: p. 38-47.
3. *Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. North American Symptomatic Carotid Endarterectomy Trial Collaborators.* N Engl J Med, 1991. **325**(7): p. 445-53.
4. Poredos, P., *Intima-media thickness: indicator of cardiovascular risk and measure of the extent of atherosclerosis.* Vasc Med, 2004. **9**(1): p. 46-54.
5. Rothwell, P.M. and C.P. Warlow, *Prediction of benefit from carotid endarterectomy in individual patients: a risk-modelling study. European Carotid Surgery Trialists' Collaborative Group.* Lancet, 1999. **353**(9170): p. 2105-10.
6. Fisher, M., et al., *The NASCET-ACAS plaque project. North American Symptomatic Carotid Endarterectomy Trial. Asymptomatic Carotid Atherosclerosis Study.* Stroke, 1993. **24**(12 Suppl): p. I24-5; discussion I31-2.
7. Naqvi, T.Z., *Ultrasound vascular screening for cardiovascular risk assessment. Why, when and how?* Minerva Cardioangiol, 2006. **54**(1): p. 53-67.
8. Roquer, J., et al., *Endothelial dysfunction, vascular disease and stroke: the ARTICO study.* Cerebrovasc Dis, 2009. **27 Suppl 1**: p. 25-37.
9. Rothwell, P.M., et al., *Prognostic value and reproducibility of measurements of carotid stenosis. A comparison of three methods on 1001 angiograms. European Carotid Surgery Trialists' Collaborative Group.* Stroke, 1994. **25**(12): p. 2440-4.

10. Schargrodsky, H., et al., *CARMELA: assessment of cardiovascular risk in seven Latin American cities.* Am J Med, 2008. **121**(1): p. 58-65.
11. Touboul, P.J., et al., *Carotid artery intima media thickness, plaque and Framingham cardiovascular score in Asia, Africa/Middle East and Latin America: the PARC-AALA study.* Int J Cardiovasc Imaging, 2007. **23**(5): p. 557-67.
12. Molinari, F., G. Zeng, and J.S. Suri, *A state of the art review on intima-media thickness (IMT) measurement and wall segmentation techniques for carotid ultrasound.* Computer Methods and Programs in Biomedicine, 2010. **100**: p. 201-221.
13. de Groot, E., et al., *Measurement of carotid intima-media thickness to assess progression and regression of atherosclerosis.* Nat Clin Pract Cardiovasc Med, 2008. **5**(5): p. 280-8.
14. Pignoli, P. and T. Longo, *Evaluation of atherosclerosis with B-mode ultrasound imaging.* J Nucl Med Allied Sci, 1988. **32**(3): p. 166-73.
15. Molinari, F., G. Zeng, and J.S. Suri, *A state of the art review on intima-media thickness (IMT) measurement and wall segmentation techniques for carotid ultrasound.* Computer Methods and Programs in Biomedicine, 2010. **100**(3): p. 201-221.
16. Faita, F., et al., *Real-time measurement system for evaluation of the carotid intima-media thickness with a robust edge operator.* J Ultrasound Med, 2008. **27**(9): p. 1353-61.
17. Molinari, F., et al., *CALSFOAM - completed automated local statistics based first order absolute moment" for carotid wall recognition, segmentation and IMT measurement: validation and bench-marking on a 300 patient database.* Int Angiol. **30**(3): p. 227-41.
18. Molinari, F., et al., *CAUDLES-EF: Carotid Automated Ultrasound Double Line Extraction System Using Edge Flow.* Journal of Digital Imaging.
19. Molinari, F., et al., *Completely automated robust edge snapper for carotid ultrasound IMT measurement on a multi-institutional database of 300 images.* Med Biol Eng Comput, (in press).
20. Molinari, F., G. Zeng, and J. Suri, *Inter-Greedy Technique for Fusion of Different Segmentation Strategies Leading to High-Performance Carotid IMT Measurement in Ultrasound Images.* Journal of Medical Systems, 2010: p. 1-15.
21. Molinari, F., G. Zeng, and J.S. Suri, *Carotid wall segmentation and IMT measurement in longitudinal ultrasound images using morphological approach*, in *2010 IEEE International Symposium on Biomedical Imaging: from Nano to Macro*2010: Amsterdam, The Netherlands.
22. Amadasun, M. and R. King, *Textural features corresponding to textural properties.* IEEE Transactions on Systems, Man and Cybernetics, 1989. **19**(5): p. 1264-1273.
23. Tan, J.H., et al., *Study of normal ocular thermogram using textural parameters.* Infrared Physics and Technology, 2010. **53**(2): p. 120-126.

24. Conners, R.W. and C.A. Harlow, *A Theoretical Comparison of Texture Algorithms.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1980. **PAMI-2**(3): p. 204-222.
25. Molinari, F., et al., *Completely automated multiresolution edge snapper-A new technique for an accurate carotid ultrasound IMT measurement: Clinical validation and benchmarking on a multi-institutional database.* IEEE Transactions on Image Processing, 2012. **21**(3): p. 1211-1222.
26. Suri, J.S., R.M. Haralick, and F.H. Sheehan, *Greedy Algorithm for Error Correction in Automatically Produced Boundaries from Low Contrast Ventriculograms.* Pattern Analysis & Applications, 2000. **3**(1): p. 39-60.
27. Loizou, C.P., et al., *Snakes based segmentation of the common carotid artery intima media.* Medical & biological engineering & computing, 2007. **45**(1): p. 35-49.
28. Bland, J.M. and D.G. Altman, *Statistical methods for assessing agreement between two methods of clinical measurement.* Lancet, 1986. **1**(8476): p. 307-310.
29. Rosati, S., F. Molinari, and G. Balestra, *Rough Set Based Approach for IMT Automatic Estimation.* International Journal of Bioelectromagnetism, 2012. **14**(4): p. 211-216.
30. Rosati, S., F. Molinari, and G. Balestra. *Feature selection applied to ultrasound carotid images segmentation*. in *Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE*. 2011.

Chapter 7: Fuzzy-Snake System for Carotid Segmentation

7.1 Introduction

In Chapter 6 we proposed a segmentation tool based on pixel classification (*PxCl*), that is able to automatically trace the LI and MA interfaces on ultrasound carotid images. Although its results were encouraging with respect to the classical semi-automatic segmentation techniques, further improvements can be achieved. The aim of this work is to propose a new method for the automatic segmentation based on the combination of a dual-snake model with the basics of the Fuzzy Logic theory.

Snakes, also called *active parametric contours*, are one of the most used techniques for image. They are deformable models evolving under the action of different forces, until equilibrium among these forces is reached. Usually, two forces are considered: an *internal* force that drives the shape of the snake, and an *external* force, which drives the snake towards the image intensity discontinuities. Snakes were commonly used for image segmentation, and in particular for LI and MA delineation and IMT measurement [1-3], with very satisfactory results. However, this technique is not completely automatic, as it requires the initialization of the contours, that means to manually define the profiles from which the evolution starts, and the setting the intensity of the forces by means of the correct choice of their coefficients. Moreover, snake shows very high sensitivity to image noise, because it is attracted by image discontinuities. Finally, snakes systems usually depict a bias in the IMT measurement that is higher

when the actual artery thickness increases (as it was showed in Chapter 6). This could constitute a limitation in the use of snake systems in the assessment of the subject's cardiovascular disease (CVD) risk.

In this technique, the snakes are automatically initialized using the LI and MA interfaces traced by *PxCl* described in Chapter 6, with the aim of further improving its performances in detecting the two interfaces of interest and thus the final IMT evaluation. Furthermore, it does not require any parameter tuning because the snake evolution is driven by a Fuzzy Inference System (FIS), taking as input the variables describing the position of the snake vertices on the image. Thus, the two most important problems related to snake implementation are definitely overcome.

7.2 Active Parametric Contours (Snakes)

From a geometrical point of view, a *snake* in a 2-dimensional space can be represented as a curve in curvilinear coordinates defined as:

$$v(s) = [x(s), y(s)]$$

where $(x, y) \in \mathbb{R}^2$ are the spatial coordinates of an image and $s \in [0,1]$ represents the parametric domain.

The points of the curve move iteratively under the action of two opposite forces, until the condition of minimal energy (balance between the forces) is reached. Usually, the global energy of a snake $E(v)$ is expressed as:

$$E(v) = E_{int}(v) + E_{ext}(v)$$

where $E_{int}(v)$ and $E_{ext}(v)$ are the internal and external energies driving the curve evolution, respectively.

The internal energy is related exclusively to the curve geometry and defines its shape and regularity. Generally, it is formulated as:

$$E_{int}(v) = \int_0^1 \{\alpha(s)|v'(s)|^2 + \beta(s)|v''(s)|^2\} ds$$

where $\alpha(s)$ and $\beta(s)$ are the coefficients related to the snake elasticity and rigidity, respectively, that avoid the curve interruption or wrapping.

The external energy is completely depending on the image characteristics, such as borders, lines, etc... , and in most applications, it is modeled using local gradients.

These two forces act on the snake configuration in opposite manner, moving each point of the curve until the equilibrium condition between the forces is reached, meaning that the snake is completely adapted to the shape or object to be segmented.

Several problems are related to the snake implementation and affect its performances [4]. Firstly, the coefficients need to be tuned in order to allow the snake to reach the equilibrium between the forces in correspondence of the interesting objects in the image. Moreover, as the snake iteratively evolves starting from an initial condition, the profile at the first iteration has to be manually defined. This means that also the number of points constituting the snake has to be chosen arbitrarily. As the perfect balance between internal and external force is hard to be reached, in most cases it is necessary to fix a maximum number of iterations for the model evolution. This number has to be selected so that the profile has enough time to reach its final shape. Finally, the snake is highly sensitive to noise that can produce gradients not related to the interesting image characteristics and therefore local minima, that can influence the final performances.

7.3 Fuzzy-Snake System

In this chapter, a completely automatic method for driving the snake evolution based on Fuzzy Logic is presented.

A dual-snake model was implemented to identify the LI and MA profiles on the carotid wall. The two snakes evolved simultaneously driven by a FIS made of 7 input variables, associated to internal and external forces, 2 output variables defining the movements of the points in the LI and MA snakes on the image, and 26 rules connecting inputs with outputs. The Mamdani method was chosen as inference technique and the centroid was used to calculate the defuzzified value. The Fuzzy Logic toolbox provided in Matlab environment was used for the FIS implementation. The schematic representation of the FIS driving the snake evolution is reported in Figure 7.1.

For the initialization of the LI and MA snakes, we used the profiles automatically traced using PxCl described in Chapter 6, moved upwards (for LI) and downwards (for MA) of 0.5 mm, as it is showed in Figure 7.2. This modification was performed because the initial profiles were too much close to the final expected interfaces and it resulted too difficult for the snakes to evolve starting from this condition. Finally, 50 iterations were set for completing the evolution.

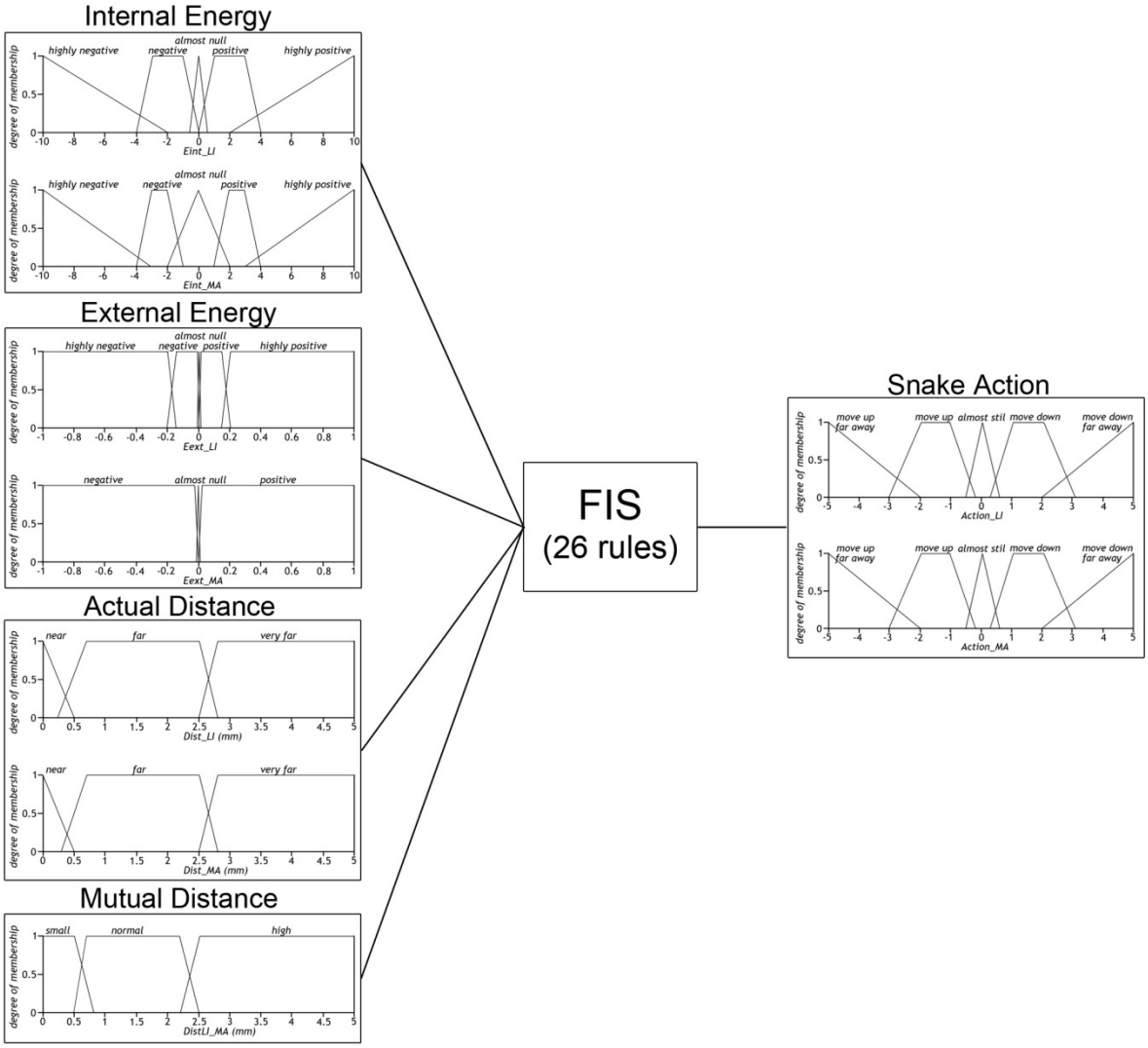

Figure 7.1 Schematic representation of the FIS driving the snake evolution. The FIS is made of 7 input variables, 2 output variables, and 26 rules connecting inputs with outputs.

7.3.1 Input Variables

Seven input variables were used for the FIS implementation: the internal energy of LI, the internal energy of MA, the external energy of LI, the external energy of MA, the actual distance of LI snake from the initial profile, the actual distance of MA snake from the initial profile, and the mutual distance between LI and MA. These parameters were calculated for all points of the snakes in each iteration, until the maximum number of iterations was reached.

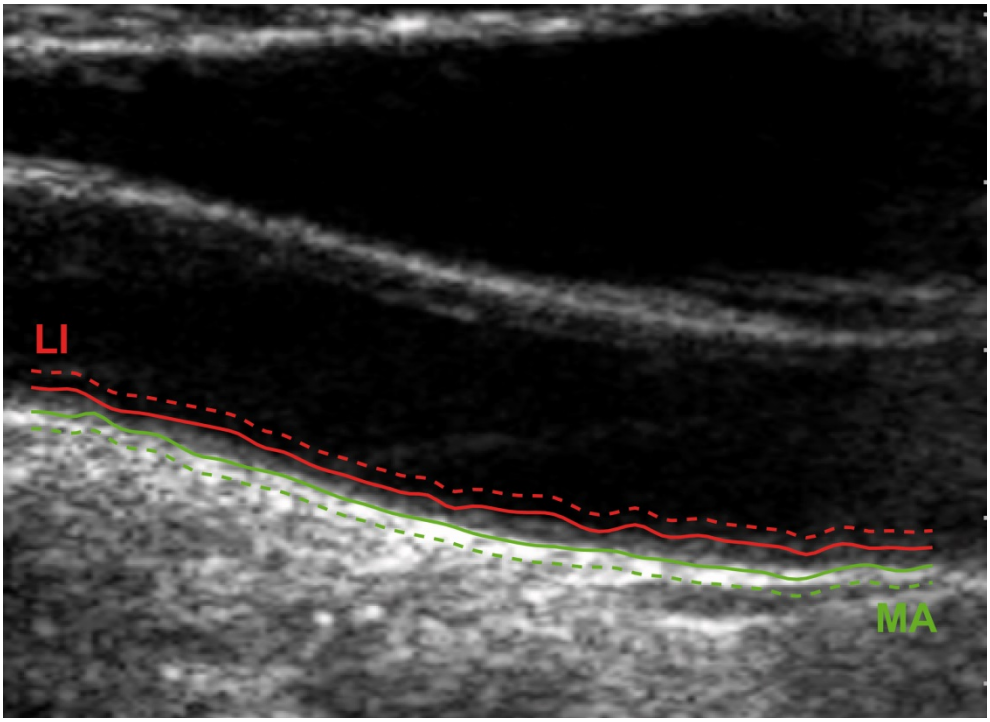

Figure 7.2 LI (red lines) and MA (green lines) interfaced automatically traced using the pixel classification tool (solid lines) and the profiles used for the snake initialization (dashed lines), moved upwards (for LI) and downwards (for MA) of 0.5 mm with respect to the first ones.

The variables were defined using Fuzzy Logic by means of a set of MFs. All MFs were constructed analyzing the values that each variable assumes when the snakes are in specific positions on the image. For this phase we used the dataset made of 120 images employed for the construction of PxCl (§Chapter 6).

In particular, three situations were taken into account, that are showed in Figure 7.3: the initial profiles, i.e. 0.5 mm up or down the LI and MA interfaces returned by PxCl (blue lines in Figure 7.3), the manually traced interfaces (red lines in Figure 7.3), and within the intima-media complex at a distance of ¼ IMT from the manual boundaries (green lines in Figure 7.3).

The input variables and the associated MFs are described in the following sections.

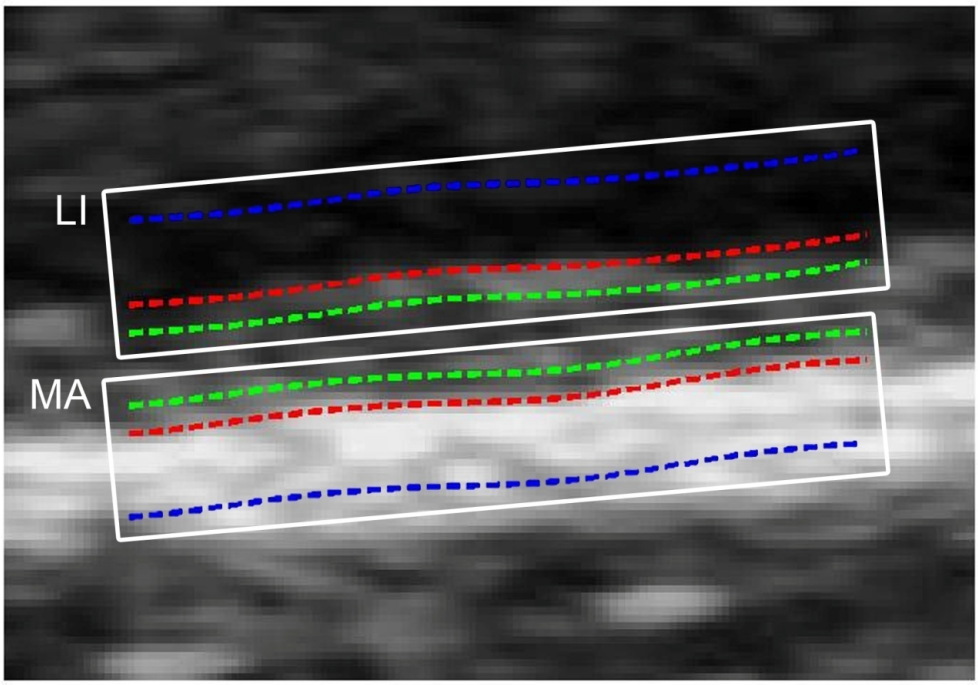

Figure 7.3 LI and MA positions analyzed for constructing the MFs of the FIS variables: in the initial snake positions, i.e. 0.5 mm up or down LI and MA manually traced (blue lines), in correspondence of the manually traced interfaces (red lines), and within the intima-media complex at a distance of ¼ IMT from the manual boundaries (green lines).

7.3.1.1 Internal Energy

A simple version of the *internal energy* was used for this fuzzy-snake implementation. It is defined for a point in the curve with coordinate y as:

$$E_{int}(y) = \frac{1}{2}(y_{left} + y_{right}) - y$$

where y_{left} and y_{right} are the ordinates of the vertices at the left and at the right of the considered point. It is related exclusively to the curve geometry and defines its shape and regularity.

The internal energy was independently defined for all points in LI (E_{int}_LI) and in MA (E_{int}_MA), and the two variables were modeled using five MFs: *highly negative, negative, almost null, positive* and *highly positive*. The MFs used for *almost null*, *highly negative* and *highly positive* were triangular-shaped, while the remaining two MFs were trapezoidal. The MFs for E_{int}_LI and E_{int}_MA are showed in Figure 7.4.

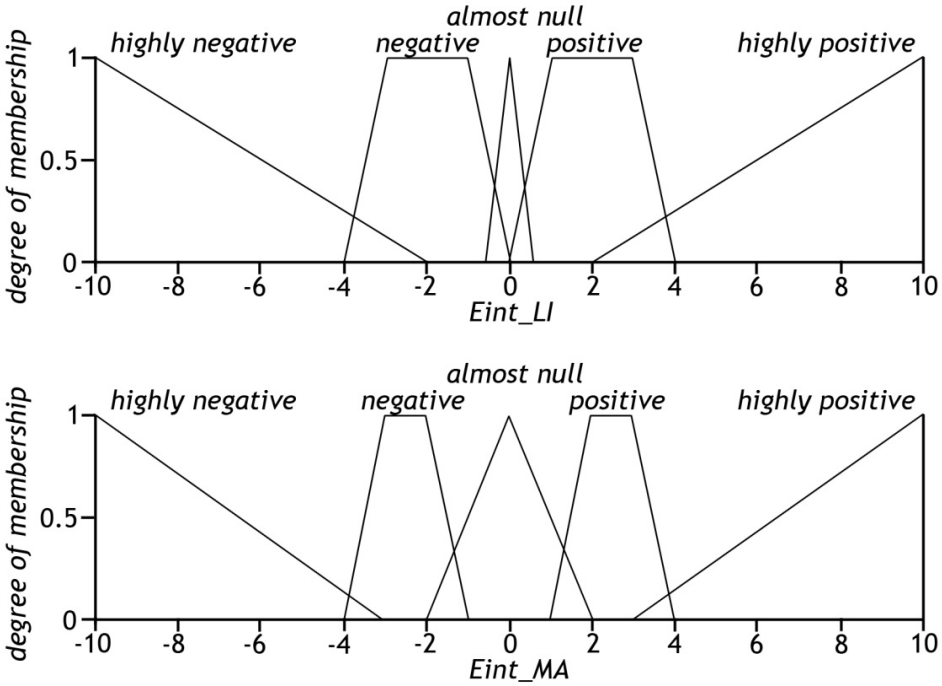

Figure 7.4 MFs for the E_{int}_LI and E_{int}_MA input variables, representing the internal energy of the points in LI and MA profiles: *highly negative*, *negative*, *almost null*, *positive* and *highly positive*.

7.3.1.2 External Energy

For calculating the *external energy*, the Sobel gradient [5] in vertical direction was used, as it highlights the vertical transitions on the image. It is formulated as:

$$h_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}.$$

Thus, the external energy for a point with coordinate y and intensity I was obtained as bidimensional convolution between the gradient and the pixel intensity as:

$$E_{ext}(y) = h_y(k,l) * I(x,y) = \sum_k \sum_l h_y(k,l)\, I(x-k, y-l).$$

As for the internal energy, also this parameter was independently defined for all points in LI (E_{ext}_LI) and in MA (E_{ext}_MA). For modeling the E_{ext}_LI we used five MFs (*highly negative*, *negative*, *almost null*, *positive*, and *highly positive*), while for E_{ext}_MA only three MFs were defined (*negative*, *almost null*, and *posi-*

tive). All MFs were modeled as trapezoid, except for *almost null* MF that was triangular-shaped. The MFs for E_{ext}_LI and E_{ext}_MA are showed in Figure 7.5.

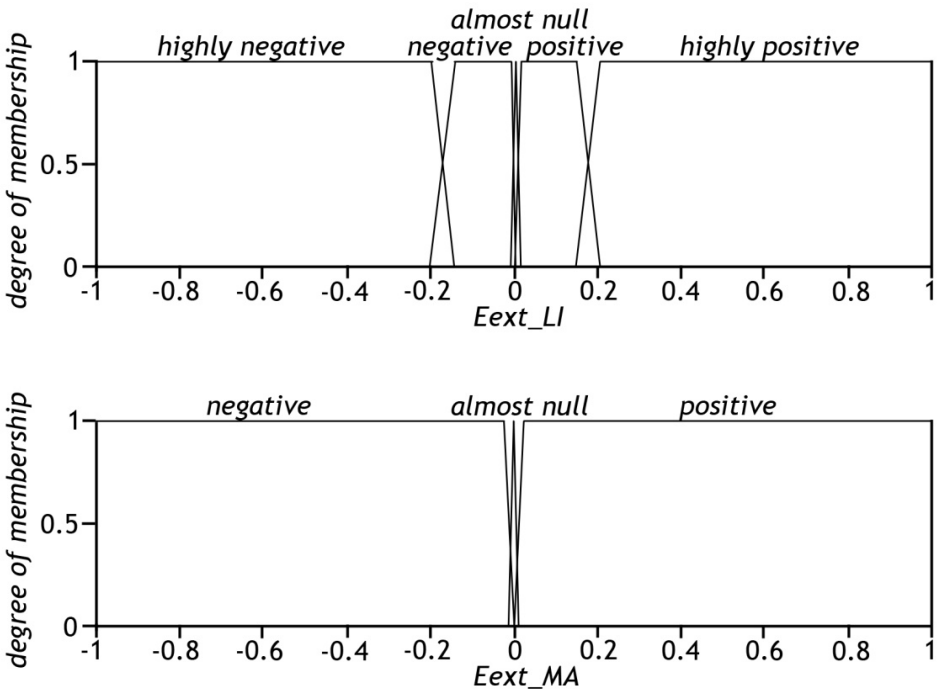

Figure 7.5 MFs for the E_{ext}_LI and E_{ext}_MA input variables, representing the external energy of the points in LI and MA profiles. For E_{ext}_LI five MFs are used (*highly negative*, *negative*, *almost null*, *positive* and *highly positive*), while for E_{ext}_MA only three MFs are implemented (*negative*, *almost null*, *positive*).

7.3.1.3 Actual Distance from the Initial Profile

This variable measures the distance (expressed in mm) of the actual snake from its initial position. It was introduced in the FIS in order to produce higher movements of the snakes when they are near to the initial profiles, and to reduce the action of the FIS when they are far enough from the starting conditions. This variable was calculated both for LI (*Dist_LI*) and MA (*Dist_MA*) and it is modeled by means of three MFs: *near* (triangular), *far* (trapezoidal) and *very far* (trapezoidal). Figure 7.6 shows the MFs associated to *Dist_LI* and *Dist_MA* variables.

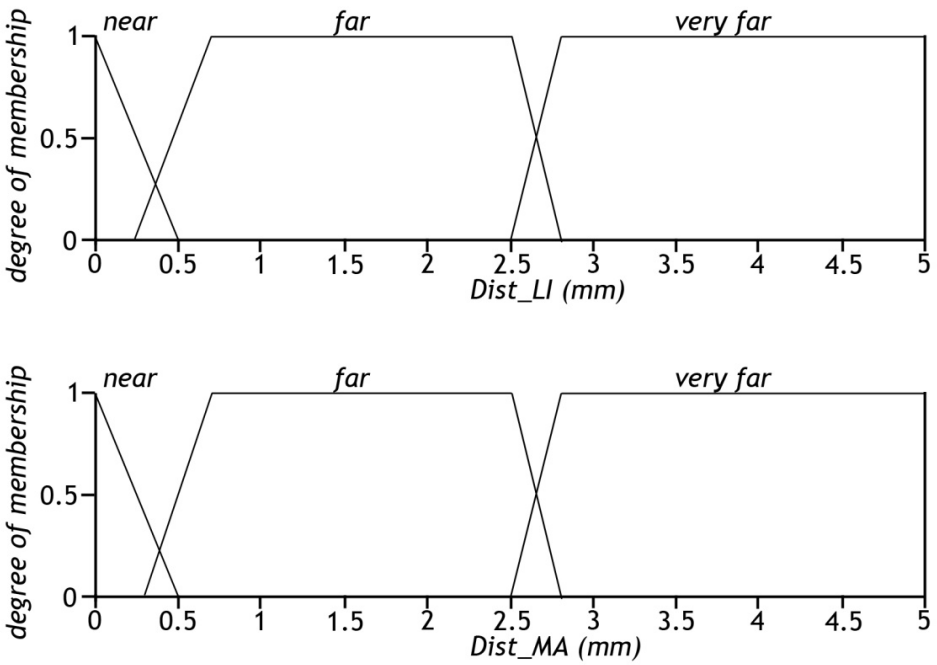

Figure 7.6 MFs for the *Dist_LI* and *Dist_MA* input variables, representing the distance (expressed in mm) of the actual snake from its initial position for both LI and MA profiles: *near*, *far*, and *very far*.

7.3.1.4 Mutual Distance

This parameter represents the distance, expressed in mm, between the actual position of LI and MA snakes and avoids the snakes to be too far or too close to each other. In this case, three trapezoidal MFs were used, *small*, *normal* and *high*, according to anatomical considerations about the admissible values for IMT. The MFs associated to *DistLI_MA* variable are depicted in Figure 7.7.

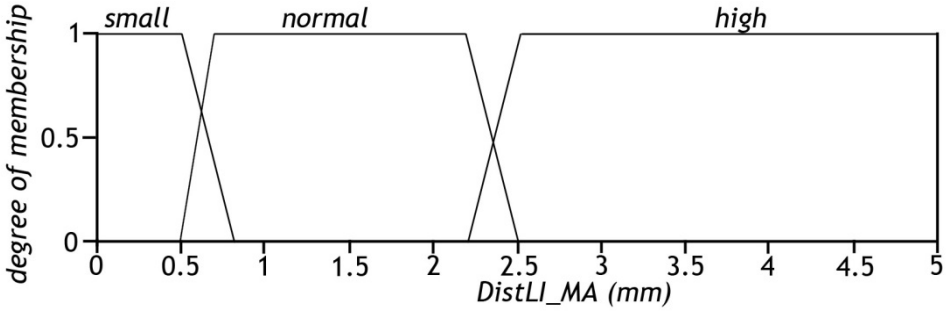

Figure 7.7 MFs for the *DistLI_MA* input variable, representing the distance (expressed in mm) between the actual position of LI and MA snakes: *small*, *normal*, and *high*.

7.3.2 Output Variables

Two output variables were defined in the FIS, representing the movement that the points in LI (*Action_LI*) and MA (*Action_MA*) snakes have to accomplish, in terms of pixels, in each iteration.

Due to the conformation of the interesting profiles and the direction of movements expected for the snakes, that are mainly vertical, only the y coordinates of the points in the two profiles is updated in each the iteration, according to the following equation:

$$y_{new} = y + y_{mov}$$

where y_{new} and y are the new and the actual coordinates of the considered points respectively, and y_{mov} indicates the movement of the points (in pixel) defined by the FIS outputs.

Five MFs were implemented to model this concept: *move up*, *move up far away*, *almost still*, *move down*, and *move down far away*. The middle and the two extreme MFs were triangular-shaped while the remaining two MFs were trapezoidal. Similarly to the input definition, also the modeling of the output variables was performed based on the analysis of the three snake positions described in section 7.3.1.

The representation of the two output variables is reported in Figure 7.8.

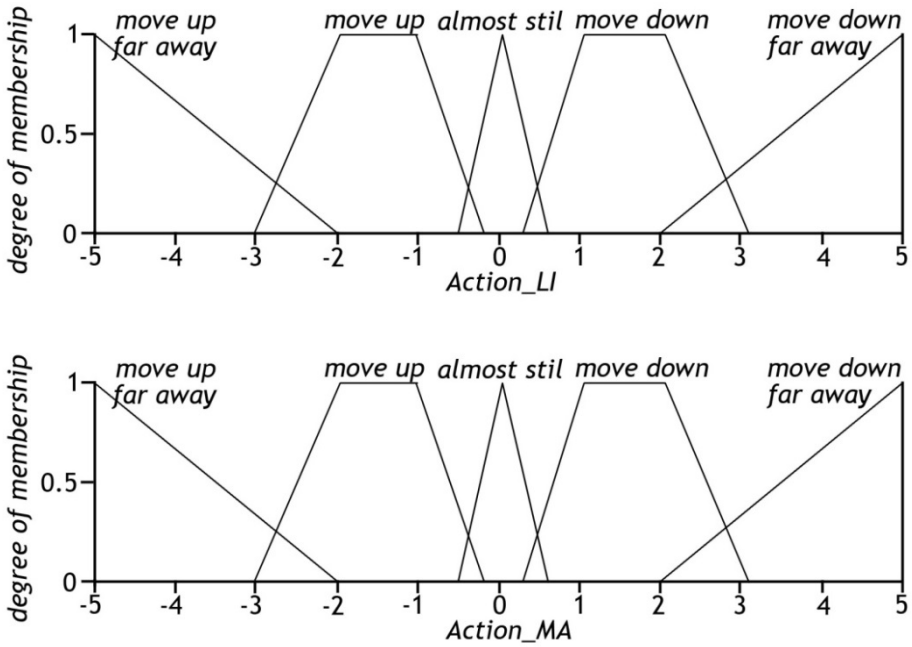

Figure 7.8 MFs for the output variables *Action_LI* and *Action_MA*, representing the action that the points in LI and MA snakes have to accomplish: *move up far away*, *move up*, *almost still, move down* and *move down far away*.

7.3.3 Rules Definition

In order to connect inputs with outputs, a set of 26 rules were implemented based on anatomical considerations, domain knowledge and expertise, and the analysis of the behavior of the two snakes during their evolution. The dataset of 120 images previously used for the MF definition was taken into account also for this phase.

Fifteen rules were defined for driving the movement of the LI snake, nine rules lead the MA profile and the remaining 2 rules define the action of LI and MA profiles at the same time. Each rule was associated to a specific weight, that could be 0.5 or 1, according to its importance for the snake movement.

The list of the 26 rules with their associated weight (reported in brackets) is given below:

1. If (E_{int}_*LI* is *highly negative*) then (*Action_LI* is *move up far away*) (1)
2. If (E_{int}_*LI* is *highly positive*) then *Action_LI* is *move down far away*) (1)
3. If (E_{int}_*LI* is *negative*) then (*Action_LI* is *move up*) (1)
4. If (E_{int}_*LI* is *positive*) then (*Action_LI* is *move down*) (1)
5. If (E_{int}_*MA* is *highly negative*) then (*Action_MA* is *move up far away*) (1)
6. If (E_{int}_*MA* is *highly positive*) then (*Action_MA* is *move down far away*) (1)
7. If (E_{int}_*MA* is *negative*) then (*Action_MA* is *move up*) (1)
8. If (E_{int}_*MA* is *positive*) then (*Action_MA* is *move down*) (1)
9. If (*DistLI_MA* is *small*) then (*Action_LI* is *move up*) and (*Action_MA* is *move down*) (1)
10. If (*DistLI_MA* is *high*) then (*Action_LI* is *move down*) and (*Action_MA* is *move up*) (1)
11. If (*Dist_MA* is *far*) then (*Action_MA* is *almost still*) (1)
12. If (*Dist_MA* is *very far*) then (*Action_MA* is *almost still*) (1)
13. If (*Dist_MA* is *near*) then (*Action_MA* is *move up)* (1)
14. If (*Dist_LI* is *far*) then (*Action_LI* is *almost still*) (1)
15. If (*Dist_LI* is *very far*) then (*Action_LI* is *almost still*) (1)
16. If (*Dist_LI* is *near*) then (*Action_LI* is *move down*) (0.5)
17. If (E_{ext}_*LI* is *almost null*) and (*DistLI_MA* is *normal*) and (*Dist_LI* is *near*) then (*Action_LI* is *move down*) (0.5)
18. If (E_{ext}_*LI* is NOT *almost null*) and (*DistLI_MA* is *normal*) and (*Dist_LI* is *near*) then (*Action_LI* is *almost still*) (1)
19. If (E_{ext}_*LI* is NOT *almost null*) and (*DistLI_MA* is *normal*) and (*Dist_LI* is NOT *near*) then (*Action_LI* is *almost still)* (0.5)
20. If (E_{int}_*LI* is *almost null*) and (*DistLI_MA* is *normal*) and (*Dist_LI* is *far*) then (*Action_LI* is *almost still*) (1)

21. If (E_{ext}_*LI* is *positive*) then (*Action_LI* is *almost still*) (0.5)
22. If (E_{ext}_*LI* is *highly positive*) and (*Dist_LI* is *near*) then (*Action_LI* is *move down*) (0.5)
23. If (E_{ext}_*LI* is *negative*) and (*Dist_LI* is *near*) then (*Action_LI* is *almost still*) (1)
24. If (E_{ext}_*LI* is *highly negative*) and (*Dist_LI* is *near*) then (*Action_LI* is *move down*) (0.5)
25. If (E_{ext}_*MA* is *almost null*) and (*DistLI_MA* is *normal*) then (*Action_MA* is *almost still*) (1)
26. If (E_{ext}_*MA* is *positive*) and (*Dist_MA* is *far*) then (*Action_MA* is *almost still*) (1)

7.4 Results

The Fuzzy-Snake system was tested on the database of 180 images used in for the validation of the PxCl (§Chapter 6). This database comprises 120 images used for the system construction, 50 images randomly selected within the available database of images and not employed for the FIS design and 10 images with plaques. The automatic LI and MA profiles obtained with the classification approach, moved of 0.5 mm up and down respectively, were considered as starting interfaces for the snakes' evolution.

Also in this case, the segmentation performed by a neurosonographer and an expert vascular radiologist were considered as Ground Truth (GT) for the system performance evaluation. The first expert will be indicated as GT1, the second as GT2. Moreover, the performances were compared also with the classification method described in Chapter 6 and with the basic version of the snakes (Classical Snake – CS) proposed by Loizou *et al.* [6], requiring the manual initialization of the two profiles.

In Figure 7.9 three examples of LI (red lines) and MA (green lines) interfaces identified by the Fuzzy-Snake (solid lines) and manually traced by GT1 (dashed lines) are showed. In Figure 7.9 panel A it is reported an image with high quality and contrast, Figure 7.9 panel B shows an image with noisy lumen, and in Figure 7.9 panel C it is showed an image with inclined vessel.

The LI and MA profiles automatically traced by the Fuzzy-Snakes and the PxCl, and the semi-automatic segmentation obtained by CS were compared with GT1 and GT2.

The distances (using the PDM metric [7]) between automatic and manual interfaces are showed in Table 7.I, in terms of mean value and standard deviation. Moreover, in the same table the results obtained with PxCl and CS are reported as comparison.

Analyzing the table it can be noticed that, using the Fuzzy-Snake tool, the average LI segmentation error is 0.18 mm and 0.11 mm with respect to GT1 and GT2 respectively. These values are lower than the ones obtained with PxCl (i.e. 0.21 mm), meaning that the initial LI profiles were actually improved. However, the MA error for the fuzzy-snake is about 0.15 mm, that is equal to the PxCl one and slightly higher than the inter-reader profile distance, which is 0.13 mm. Moreover, it emerges a clear predominance of the fuzzy-snake tool with respect to CS, giving errors higher than 0.3 mm.

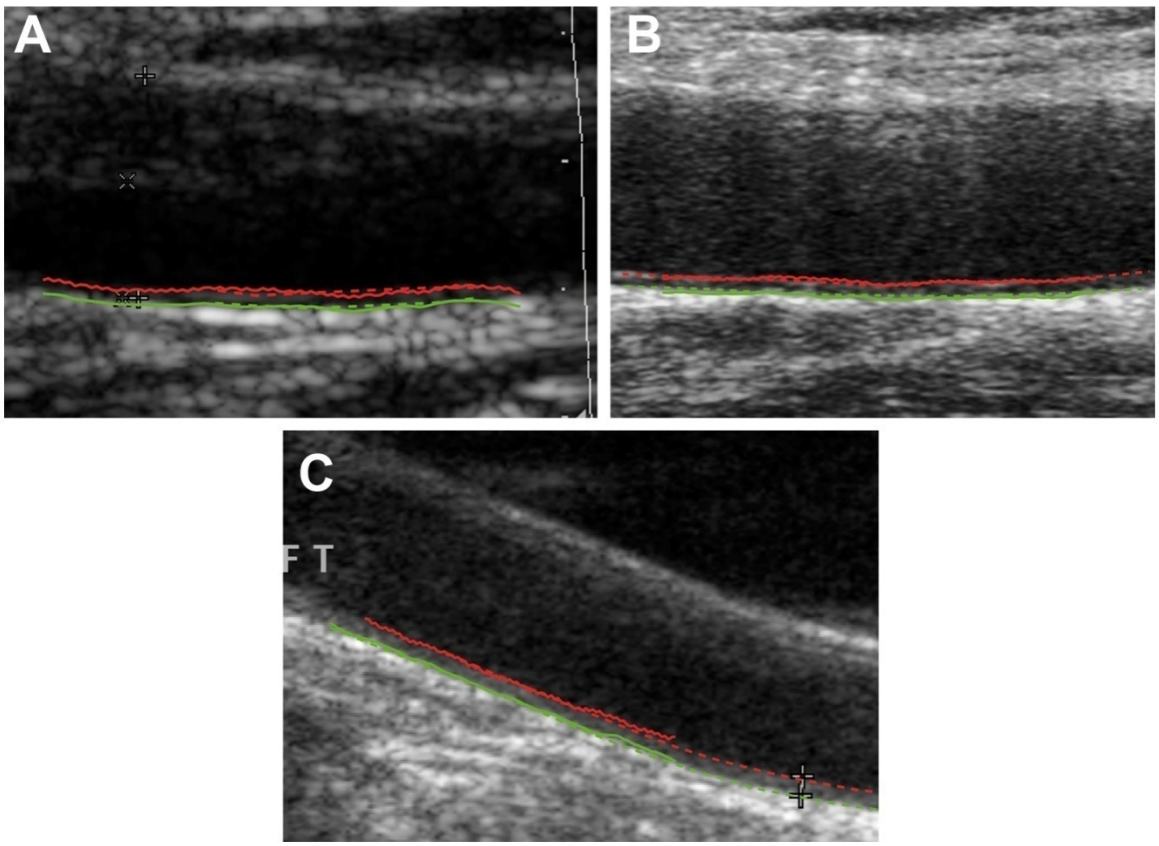

Figure 7.9 Three examples of LI (red lines) and MA (green lines) interfaces traced by the Fuzzy-Snake (solid lines) and manually traced by GT1 (dashed lines). In panel A it is reported an image with good quality and contrast, panel B shows an image with noisy lumen, and in panel C an image with inclined vessel is showed. In yellow it is traced the far adventitia boundary.

Table 7.II reports the bias and the absolute error comparing the IMT evaluation obtained with the fuzzy-snake, PxCl and CS, with GT1 and GT2. Analyzing the IMT errors, it can be observed that the IMT measurement biases of the fuzzy-snake (0.02 mm and 0.03 mm with respect to GT1 and GT2) result higher than the error between the two experts (0.01 mm) and lower that the one obtained with PxCl (about 0.5 mm). Moreover, while the profiles obtained with PxCl tended to underestimate the IMT values, the new fuzzy-snake tool tends to trace LI and MA profiles in a more conservative way, that is overestimating the IMT and, thus, the cardiovascular risk, similarly to CS.

Table 7.I Performance evaluation comparing the LI and MA interfaces obtained by the fuzzy-snake, the pixel classification method (PxCl) and the classical snakes (CS) with the two sets of human-traced profiles (GT1 and GT2), in terms of distance between the profiles measured with the PDM metric (mean value ± standard deviation).

		vs. GT1	vs. GT2
GT1	**LI**	-	(0.130±0.060) mm
	MA	-	(0.134±0.100) mm
CS	**LI**	(0.343±0.214) mm	(0.321±0.207) mm
	MA	(0.335±0.264) mm	(0.323±0.259) mm
PxCl	**LI**	(0.214±0.117) mm	(0.211±0.121) mm
	MA	(0.142±0.083) mm	(0.154±0.095) mm
Fuzzy-Snake	**LI**	(0.181±0.104) mm	(0.171±0.103) mm
	MA	(0.141±0.079) mm	(0.158±0.094) mm

Table 7.II Performance evaluation comparing the IMT values obtained by the fuzzy-snake, the pixel classification method (PxCl) and the classical snake (CS) with the two sets of human-traced profiles (GT1 and GT2) in terms of bias and absolute error (mean value ± standard deviation).

		vs. GT1	vs. GT2
GT1	**Error**	-	(0.006±0.203) mm
	|Error|	-	(0.132±0.155) mm
CS	**Error**	(0.002±0.290) mm	(0.008±0.260) mm
	|Error|	(0.206±0.203) mm	(0.194± 0.172) mm
PxCl	**Error**	(-0.052±0.227) mm	(-0.047±0.228) mm
	|Error|	(0.177±0.151) mm	(0.173±0.156) mm
Fuzzy-Snake	**Error**	(0,023±0.220) mm	(0.030±0.200) mm
	|Error|	(0.169±0.142) mm	(0.154±0.130) mm

Concerning the IMT error absolute value, it is 0.17 mm and 0.15 mm for the fuzzy-snake with respect to GT1 and GT2 respectively. These values are statistically different from the CS ones, in terms of both mean value ($p<1\%$) and variance ($p<1\%$). The increasing in the performance are also evident with respect to PxCl, especially in the comparison with GT2 (Student t-test: $p<5\%$, Fisher F-test: $p<5\%$).

Figure 7.10 shows the histograms of the error distributions in the IMT evaluation, for the whole set of images. Figure 7.10 panel A reports the bias with respect to GT1, while in Figure 7.10 panel B it is showed the error with respect to GT2. Moreover, the distributions related to CS and the two human experts are reported as comparison.

Analyzing the graphs it can be observed that the normal distribution curves obtained with the fuzzy-snakes (violet dashed lines) are almost overlapping to the manual ones (black dashed lines), in particular with respect to GT2 (Figure 7.10 panel B). The width of these curves is also clearly smaller than that related to CS (green dashed lines), as it is confirmed by the lower values of standard deviation reported in Table 7.II.

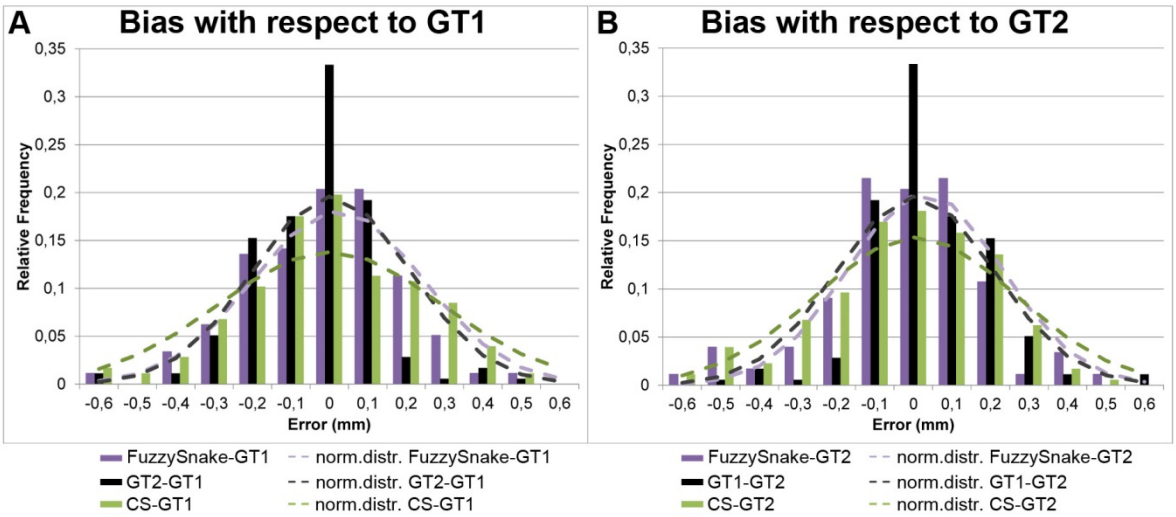

Figure 7.10 Histograms of the error distribution in the IMT evaluation for the set of 180 images. Panel A reports the bias of the fuzzy-snakes with respect to GT1, while in panel B it is showed the error with respect to GT2. Moreover, the distributions related to CS and the two human experts are reported as comparison. The dashed lines represent the normal distribution curves for each comparison.

Figure 7.11 reports the Bland–Altman plots of automatic and semi-automatic snakes compared with the two set of manual interface, in terms of IMT measurement for 180 pathologic and non-pathologic carotid vessels. In Figure 7.11 panels A and B the fuzzy-snake system is compared with GT1 and

GT2, respectively; in Figure 7.11 panels C and D the comparison between CF and the two human operators is showed.

The Bland-Altman plots show the congruence between the fuzzy-snake outcomes and the profiles traced by GT1 and GT2 (Figure 7.11 panels A and B). In fact, it can be observe that almost all points are within the dashed lines. Although the overall distribution of the points in these two cases shows a negative trend between the measurements sets, this is less evident than in the plots related to CS (Figure 7.11 panels C and D), above all for IMT values higher than 0.85 mm (red areas).

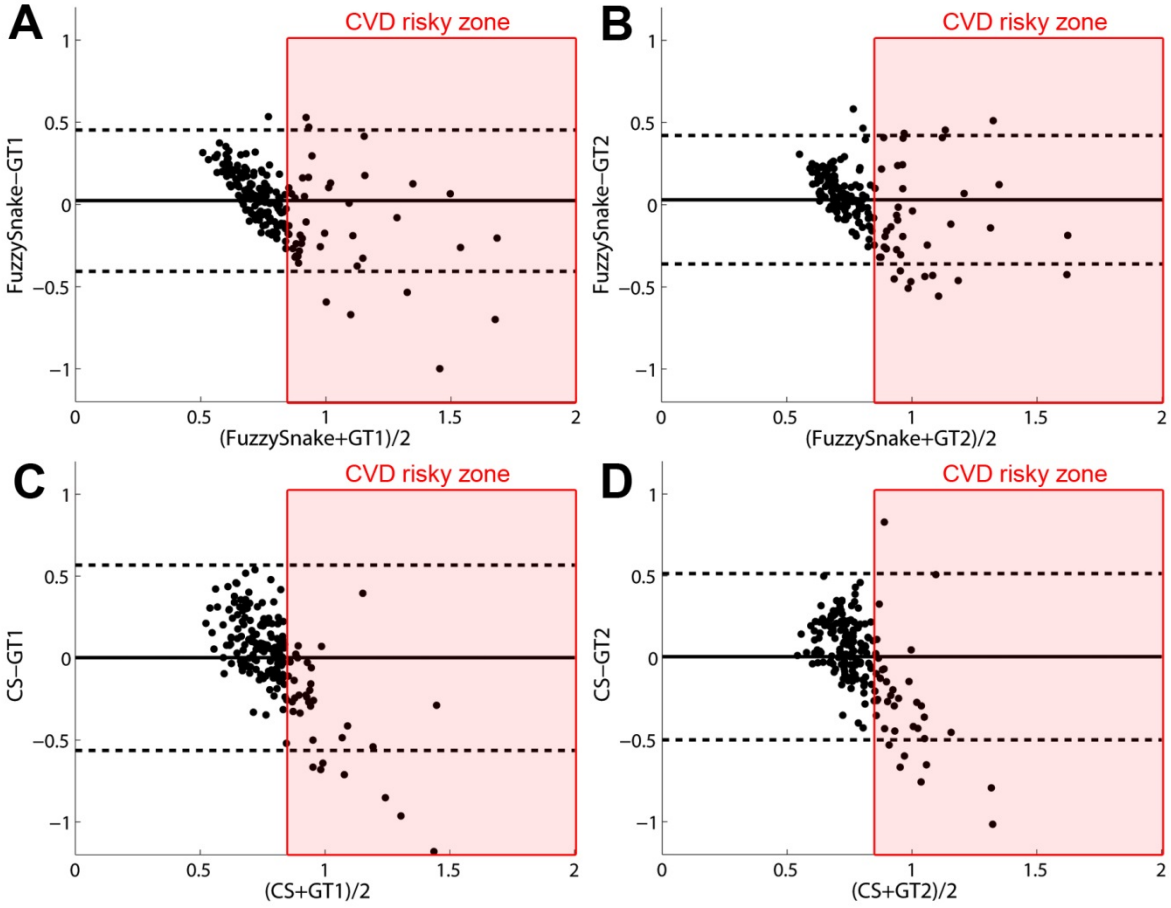

Figure 7.11 Bland–Altman plots of automatic and semi-automatic snakes compared with the two set of manual interface, in terms of IMT measurement for 180 pathologic and non-pathologic carotid vessels. In panels A and B the fuzzy-snake system is compared with GT1 and GT2, respectively; in panels C and D the comparisons between CF and the two human operators are showed. The solid lines represent the mean difference between the two sets of measurements while the dashed lines are placed in correspondence of mean ± 1.96*standard deviation. The red areas highlight the CVD risky zones (with IMT higher than 0.85 mm).

7.5 Discussion

In this work we present a new segmentation method for automatic detection of LI and MA interfaces. This technique is based on the simultaneously evolution of the two snakes under the action of a FIS. In addition to the internal and external energy commonly taken into account for standard snake implementations, also the mutual distance between the two interfaces was introduced as input parameter, in order to avoid the snake collapsing or being too far. Moreover, the distances between the actual snake positions and their initial ones were used as inputs. These last two variables, apart from pushing the snakes to move from their initial conditions, force the profiles to evolve faster during the first iterations. Consequently, the balancing among the different forces is reached sooner than for traditional snake implementations: 50 iterations are required by the fuzzy-snakes for completing their evolution while the non-fuzzy version usually needs about 200 iterations [8].

We tested our method for 180 multi-institutional, multi-ethnic, multi-scanner, and multi-operator images, both with normal carotid wall and with plaques. The results were compared with PxCl, CS and two sets of manually traced profiles taken as ground truth.

With respect to PxCl profiles, that were used in this study for initializing the two snakes, a clear improvements are obtained with fuzzy-snake in terms of error in tracing the LI profile (Table 7.I). In fact, in this case we passed from 0.21 mm to 0.18 mm of mean LI distance from the manual interfaces. This is very important for clinical applications on pathological images, in which the LI interface could be altered from the presence of plaques and thus it is needed to tracing it as correctly as possible. However, no improvements are observable with respect to PxCl for the MA profile, that was already enough accurate.

The more precise detection of LI consequently affects the performances in the IMT estimation. Although the fuzzy-snake results still more biased for the mean IMT value than CS, its absolute error and standard deviation are now evidently lower than those obtained with the semi-automatic technique (Table 7.II), meaning that the use of a FIS for driving the snake evolution allows obtaining more precise and reproducible measurements.

As it emerges from the Bland-Altman plots, overall the CS measurements of IMT are all lower than the ground truth for carotid thickness higher than 0.85 mm (Figure 7.11 panels C and D). Conversely, although the fuzzy-snake performance shows a negative trend with respect to GT1 and GT2 (Figure 7.11

panels A and B), this is not so evident for higher values of IMT, for which we obtained both positive and negative errors.

The underestimation of the actual IMT value for thicker walls by snake systems is a common problem. Underestimation is often caused by the mutual energy that constrains the position of the LI snake with respect to the MA snake. For example, in the dual-snake system proposed by Molinari et al. [9], an external force acted on the snakes in order to prevent them from collapsing and from being too distant. Hence, there was a parameter that calibrated this force. As mentioned in the introduction, the optimization of the snake parameters is not a trivial task. This problem is crucial when snakes are used to assess the subject's CVD risk. In fact, advanced atherosclerosis leads to higher IMT values, which could be underestimated by a traditional snake-based algorithm (as in Figure 7.11 panels C and D). This might lead to a systematical underestimation of the CVD risk.

Our novel automated fuzzy-snake system does not systematically underestimate IMT in case of advanced atherosclerosis. The adoption of fuzzy logic allowed for an easy optimization of the snake evolution. The fuzzyfication of the mutual distance between the LI/MA snakes does not prevent the snake from correctly measuring higher values of IMT.

7.6 Conclusions

In this chapter, we described a new tool for improving the segmentation of ultrasound carotid images obtained with PxCl reported in Chapter 6. This system represents the evolution of the traditional snake approach by using Fuzzy-Logic. In particular, a FIS was implemented for driving the evolution a dual-snake system that moves on the ultrasound image until the LI and MA interfaces are recognized.

The fuzzy-snake evolution is based on seven input parameters calculated for each point in the profiles. Besides the classical internal and external forces, we introduced also the contribution of the mutual distance between the two snakes, avoiding the snakes to be too close or too far to each other.

This new tool overcomes the main limitations of the classical snakes because it don't need the user interaction for initializing the snakes and for choosing the correct parameters, as the FIS is able to automatically balance different forces contributions. Moreover, the fuzzy system is designed so that the snake equilibrium is reached faster (within 50 iterations) than in the classical version, thanks to the variables related to the actual distance of the snake from its start-

ing point. Finally, as it is based on the classical snake algorithm, it is able to work with different kinds of images, in terms of vessel morphologies and levels of noise.

The proposed tool effectively represents an improvement in the segmentation performances with respect to the PxCl, especially for the LI profile that results more similar to the human-traced one. This increased accuracy of the fuzzy-snake in the delineation of the interfaces makes it a suitable technique for the measurement of the intima-media thickness variability along the carotid.

When compared to a traditional snake implementation, the fuzzy-snake showed better performance. In particular, this tool is capable to correctly estimate the actual carotid wall thickness both for low risk subjects (i.e. with IMT lower than 0.85 mm) and for pathological subjects (i.e. with higher IMT value). Therefore, this model could be used as computerized aid in clinical studies for the accurate assessment of subjects' cardiovascular risk.

References

1. Molinari, F., et al., *Constrained snake vs. conventional snake for carotid ultrasound automated IMT measurements on multi-center data sets.* Ultrasonics, 2012. **52**(7): p. 949-961.
2. Xu, X., et al., *Ultrasound intima–media segmentation using Hough transform and dual snake model.* Computerized Medical Imaging and Graphics, 2012. **36**(3): p. 248-258.
3. Cheng, D.-c., et al., *Using snakes to detect the intimal and adventitial layers of the common carotid artery wall in sonographic images.* Computer Methods and Programs in Biomedicine, 2002. **67**(1): p. 27-37.
4. Molinari, F., G. Zeng, and J.S. Suri, *A state of the art review on intima-media thickness (IMT) measurement and wall segmentation techniques for carotid ultrasound.* Computer Methods and Programs in Biomedicine, 2010. **100**(3): p. 201-221.
5. Meenalosini, S., J. Janet, and E. Kannan, *A novel approach in malignancy detection of computer aided diagnosis.* American Journal of Applied Sciences, 2012. **9**(7): p. 1020-1029.
6. Loizou, C.P., et al., *Snakes based segmentation of the common carotid artery intima media.* Medical & biological engineering & computing, 2007. **45**(1): p. 35-49.
7. Suri, J.S., R.M. Haralick, and F.H. Sheehan, *Greedy Algorithm for Error Correction in Automatically Produced Boundaries from Low Contrast Ventriculograms.* Pattern Analysis & Applications, 2000. **3**(1): p. 39-60.
8. Molinari, F., et al., *Constrained snake vs. conventional snake for carotid ultrasound automated IMT measurements on multi-center data sets.* Ultrasonics, 2012. **52**(7): p. 949-61.

9. Molinari, F., et al., *Fully automated dual-snake formulation for carotid intima-media thickness measurement: A new approach.* Journal of Ultrasound in Medicine, 2012. **31**(7): p. 1123-1136.

Chapter 8: Conclusions

The purpose of this thesis is to propose a new methodology for analyzing biomedical data, that can be used both for extracting useful knowledge from real datasets and for improving signal and image processing techniques. Starting from the whole KDD process applied for data interpretation, we focused on the important step of FS, highlighting limitations and advantages of different methods proposed in literature.

First of all, it is important to clarify that KDD provides only a guideline for automatically extracting new knowledge from large and multidimensional datasets, that cannot be manually analyzed. In particular, KDD is a non-trivial process involving several phases that must be properly performed and adapted to the context in order to extract information, models, structures or, in general, high-level descriptions from data. Unfortunately, in the last years the research in this field was mostly related to the aspects of DM, but this step is useless without a correct choice of data and an improvement in their quality.

Moreover, the application of such techniques in the biomedical field is very poor in literature, essentially because of several difficulties in managing with this kind of data. Medical datasets, in fact, are often characterized by a small number of objects (usually patients involved in the study) associated with a very large number of variables extracted from medical analysis. Other problems concern the imbalance in the represented classes and the incompleteness and uncertainty affecting the data, derived from measurement or human errors. All these aspects make the analysis of biomedical datasets very challenging, requiring the employment of advanced techniques for managing incomplete and hidden information.

The first two works proposed in this thesis dealt with the extraction of knowledge from two small and imbalanced medical datasets. In these applications, related to the characterization of cerebral hemodynamics in migraineurs and of the muscle oxygenation pattern in diabetic subjects, different FS tools were applied and compared. In particular, we analyzed a classical statistical method (ANOVA) and advanced FS tools based on RST.

In the analyzed scenarios, the statistical methods clearly showed their limitation in dealing with real medical data. In particular, they were not able to correctly identify those aspects really relevant in the datasets for the proper classification of the subjects. Several reasons can be at the base of this behavior. In fact, such kind of tools requires linearity assumption between variables and classes, that is not always verified for medical parameters.

Furthermore, the advanced methods based on RST selected a subset of features related to the most relevant physiological aspects characterizing the pathological subjects.

These non-linear tools based on RST proved their effectiveness also in handling large amount of data, as in the case of the pixel classifier proposed in Chapter 6. In this case, the FS was applied, not for the direct discovery of knowledge, but for exploiting the information encapsulated in the dataset for the development of a new system for the automatic segmentation of the carotid wall. This is an example of how the proposed methodology can be also used as intermediate step for further phases of processing and interpretation of signals and images.

As it emerges from all applications described in this thesis, one of the most important aspects of the proposed approach for managing biomedical data and signals lies in its versatility and capability to adapt to different data, contexts and purposes. In all cases, in fact, we obtained very good results in terms of new knowledge about the domain of interest or improvements of the existent methodologies. Finally, the process described in this work for supervised applications can be easily adapted also for unsupervised purposes. In this case it is possible to substituting the supervised non-linear tools for FS with the corresponding unsupervised versions, such as the Unsupervised QuickReduct Algorithm [1], and replacing the classification methods with clustering algorithms, like k-means or SOMs (§3.1.1).

References

1. Velayutham, C. and K. Thangavel, *Unsupervised Quick Reduct Algorithm Using Rough Set Theory.* Journal of Electronic Science and Technology, 2011. **9**(3): p. 193-201.